DOCUMENT RESUME

ED 125 525 IR 003 399

AUTHOR	Yelon, Stephen L.
TITLE	Constructive Evaluation; Improving Large Scale Instructional Projects.
SPONS AGENCY	Children's Television Workshop, New York, N.Y.
PUB DATE	74
NOTE	365p.
EDRS PRICE	MF-$0.83 HC-$19.41 Plus Postage.
DESCRIPTORS	*Curriculum Evaluation; Data Analysis; Data Collection; Educational Programs; *Evaluation; *Evaluation Methods; Evaluation Needs; Failure Factors; *Formative Evaluation; *Instructional Programs; Interviews; *Program Evaluation; Site Analysis; Skill Analysis; Success Factors; Summative Evaluation; Test Construction; Testing

ABSTRACT

This text provides directors of instructional programs with an extensive overview of the evaluation process. In 25 chapters, the contents focus on the definition of evaluation, a rationale for its use, a list of tools used in the evaluation process, a delineation of the elements contained in a well-done evaluation, and some suggestions on ways that evaluations can be used to improve instructional programs. (EMH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED125525

CONSTRUCTIVE EVALUATION

Improving Large Scale Instructional Projects

by

Stephen L. Yelon

© 1974

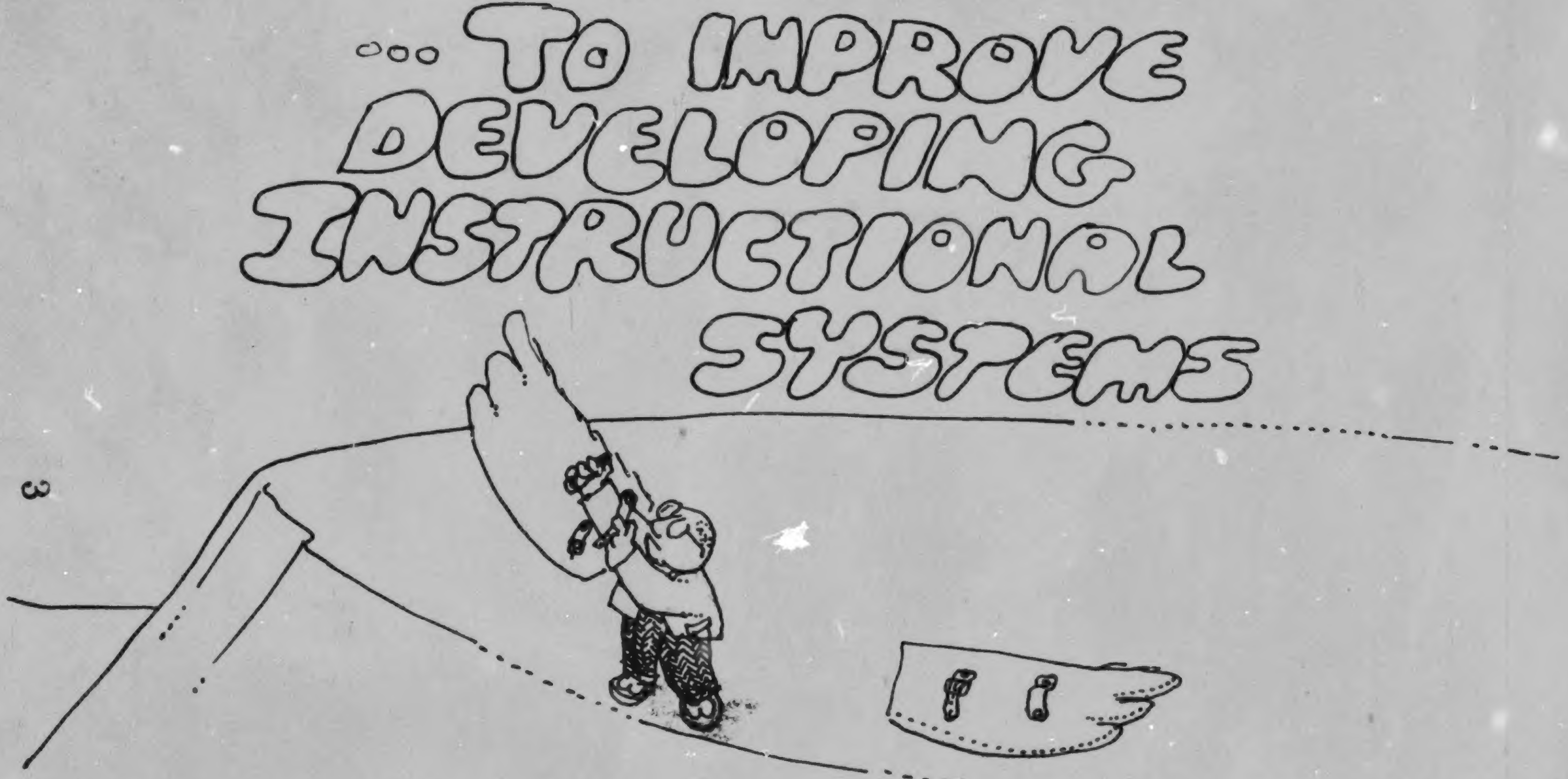
"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Stephen L. Yelon and
Children's Television Workshop
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

Stephen L. Yelon
1736 North Hayford Avenue
Lansing, Michigan, 48912

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IR003 399


...TO IMPROVE
DEVELOPING
INSTRUCTIONAL
SYSTEMS

Dedication

The process of constructive evaluation is dedicated to Daedelus, who died while trying to fly with the wax wings he and his father made; he certainly could have benefited from constructive evaluation. It's also dedicated to the makers of the Edsel, and to most educators who create their own waxed wings and Edsels every day. Every one of them must know what constructive evaluation is and how it is done.

PREFACE

The purpose of this book is to show instructional project directors, producers, and evaluators how to improve a large scale instructional system through the process of constructive evaluation. In using this method to make a course more effective, efficient, and acceptable you collect and apply data as the material is being developed.

Please note that for clarity and consistency, the messages in each chapter are addressed primarily to the director of the project and that the process of constructive evaluation can be applied to any size project, if the process is needed.

What I refer to as constructive evaluation is what some other instructional designers may call formative evaluation or developmental testing. I use the term constructive evaluation for two reasons; I want to imply that the process is positive, practical, and productive, and I want to distinguish between evaluation for developing and improving programs, methods, and materials, and other meanings, such as evaluation for diagnosing and prescribing for an individual student's learning problems.

Many techniques of constructive evaluation are described. The many combinations of procedures which are possible will help you to tailor-make your own approach. You will be able to choose those procedures most applicable to your program and you will be able to recognize a properly functioning constructive evaluation process.

The primary goal of this effort is to enable you to conduct a successful constructive evaluation, using this book as a guide. But there is no substitute for direct experience: to learn how to administer a constructive evaluation you will have to try it.

S.L.Y.

Acknowledgements

Writing a book is like balancing on top of a human pyramid. To perform well, an author must have support; he must depend on the physical, financial, and psychological assistance of many men and women. During the 1972-1973 academic year I was privileged to have such help. First, my colleagues at Michigan State University granted me a sabbatical leave and thus provided time to write. Next, the Spencer Foundation supplied funds to the Vice President for Research of the Children's Television Workshop, Edward Palmer. Then, Dr. Palmer offered me financial support, a place to work, and the opportunity to explore the general topic of constructive evaluation.

All those people who helped me in my investigation had to take time from their own pursuits, and bear the weight of what must have seemed like naive, incessant inquisition. I owe much to the production staff of the Children's Television Workshop; they taught me about the production of excellent children's television. The researchers at C.T.W. answered my questions, allowed me to observe their activities, and let me use their documents. For this I thank Trish Hayes, formerly Director of "Sesame Street" research, Vivian Horner, Director of "The Electric Company" research, and Joyce Weil, formerly director of the health show research. Some C.T.W. staff members shouldered more weight than others. I thank Lewis Bernstein, Frances Aversa, and George Beker for their continued encouragement.

To collect information about the state of the art I visited the Southwest Regional Laboratory for Educational Research and Development,

the Far West Laboratory for Educational Research and Development, Bilingual Children's Television, and Michigan State University. I appreciate the advice and the information given to me by the staff of these agencies:

Michigan State University
    Allan J. Abedor
    William Schmidt
Far West Regional Laboratory
    Nicholas Rayder
    Jerold Walker
    Richard Watkins
    Moriss Lai
    Meridith Gall
    Linda Sikorski
Bilingual Children's Television
    Lily Fillmore
    Heidi Dulay
Southwest Regional Laboratory
    Roger Scott
    James Maloney
    Richard Schutz
    Ray Briggs

I also interviewed Stephen Klein of The Center for the Study of Evaluation at the University of California at Los Angeles. I give special thanks to my friend Nicholas Rayder for introducing me to the staff of the F.W.R.L., and I am forever grateful to my friend Roger Scott for setting up my appointments at the S.W.R.L. and for his insightful, critical comments and his freedom in the sharing of ideas.

George Beker was one of my daily contacts. George Beker not only provided encouragement but converted evaluation ideas into the fine drawings in this text. Dianne Muhlfeld, Alisa Lopano, at C.T.W., and Janice Miersma at M.S.U. strained their eyes and broke their fingernails in the process of typing draft after draft of the manuscript.

Great thanks belongs to those who read and edited each draft of the manuscript. While caring for a husband, three children, and a home with her left hand, Carolyn Owens used her right hand to read, reread, and provide fine editorial criticism. My wife Frances edited each draft before it went to Carolyn and comforted me when I read Carolyn's comments.

I could not possibly have written this text without the help of all these people. Each had to be there to play their role in holding the pyramid together.

TABLE OF CONTENTS

CHAPTER I Introduction 1

CHAPTER II The Big Ball of Wax: An Overview 5

CHAPTER III Mental Gymnastics: Determining the Need for Constructive Evaluation 19

CHAPTER IV Writing the Recipe: Forming Evaluation Questions 63

CHAPTER V The Bridge: Defining Elements in an Evaluation Question 67

CHAPTER VI The Tailor's Tape and Assorted Supplies and Tools: The Elements Required for a Test of an Instructional Project 91

CHAPTER VII Tool Number One: The Review 95

CHAPTER VIII Tool Number Two: The Progress Measure 109

CHAPTER IX Tool Number Three: The Criterion Test 125

CHAPTER X Tool Number Four: The Rating Form 131

CHAPTER XI Tool Number Five: The Interview 141

CHAPTER XII The Test of a Test: Standards for Judging a Constructive Evaluation Test 149

CHAPTER XIII Supply Number One: A Prototype Unit 165

CHAPTER XIV Supply Number Two: A Sample of Students 173

CHAPTER XV Supply Number Three: A Test Site 189

CHAPTER XVI Trial for Error: Organizing and Conducting a Tryout 201

CHAPTER XVII Assembling the Puzzle: Organizing the Data 225

CHAPTER XVIII Studying the Puzzle: Analyzing the Data 239

CHAPTER XIX Detective at Work: Identifying the Strengths and Weaknesses of the Instructional Method 245

CHAPTER XX The Puzzle of Keys and Locks: Identifying the Factors Which Contribute to Success and Failure 253

CHAPTER XXI Disciplined Creativity: Extracting Design Principles 277

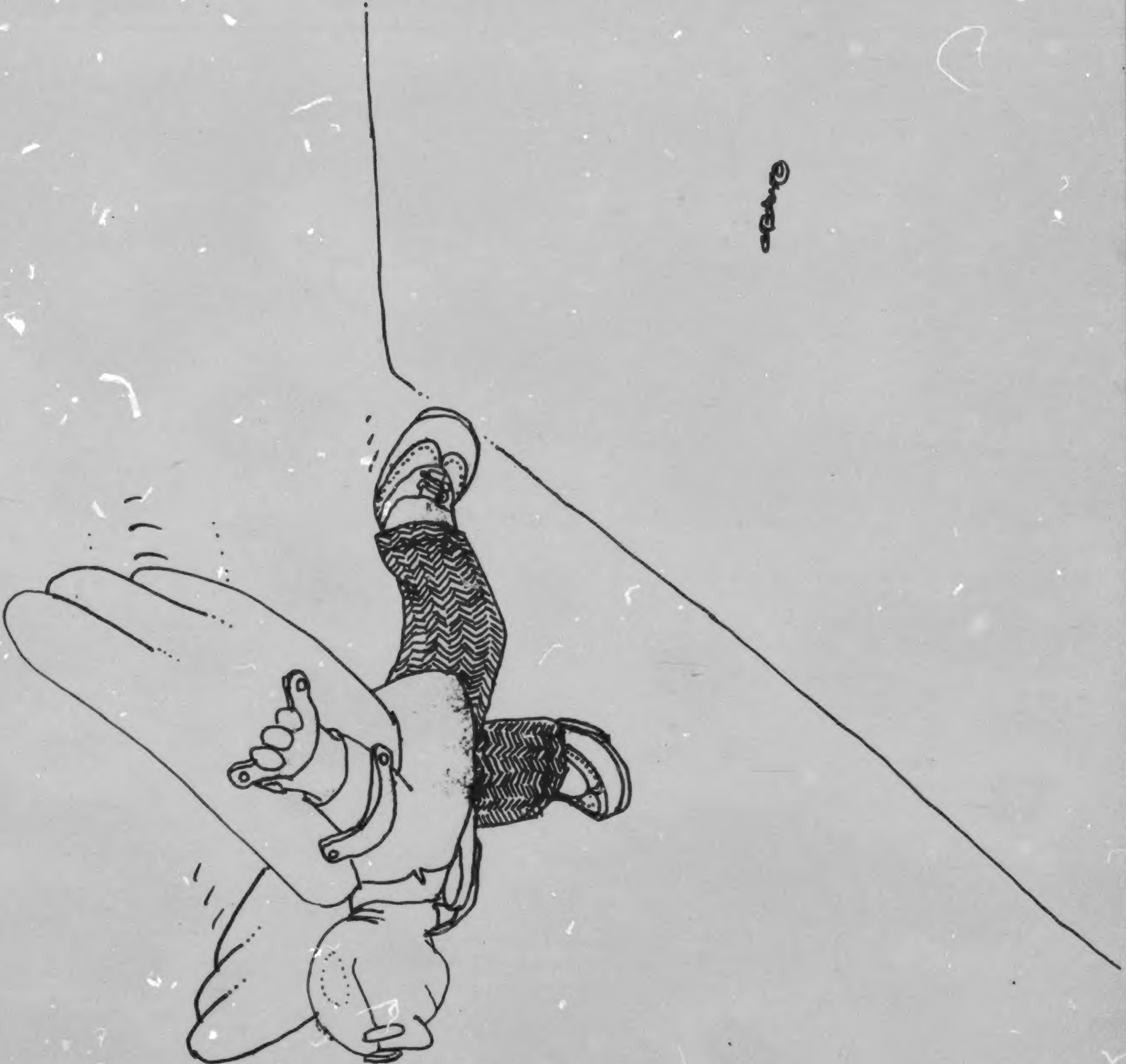
CHAPTER XXII Metamorphosis: Generating Modifications 285

CHAPTER XXIII Try, Try Again: Recycling 305

CHAPTER XXIV The News: Reporting Constructive Evaluation Results 317

CHAPTER XXV The Odd Couple: Working Toward Commitment 331

References 349



CHAPTER I

INTRODUCTION

* * * * * * * *

A Legend

In ancient times when gods walked the face of the earth, man was often reminded of his imperfect nature. In one case, the god of instruction, Pedagogio, confronted educators: "Your efforts bring mediocre results and yet you are satisfied. Your instruction is imperfect, and you make little effort to improve." The teachers exclaimed, "That is not true. Our work is good; at least, there is no evidence to the contrary." Pedagogio smiled and said, "Go and create a Great Lesson and teach it to all the people, and we shall see." The educators followed Pedagogio's bidding; and then Pedagogio collected evidence of student learning to show the teachers the results of their work.

To his great surprise, Pedagogio found results indicating some success: many students were learning. Generally, however, the data confirmed Pedagogio's pronouncements: many students were not learning. When Pedagogio showed the teachers his findings, he was surprised again, this time by their reaction. Instead of making excuses, the teachers set out to improve their instruction in order to multiply their successes and reduce their failures. When the teachers felt they had improved their lesson sufficiently, they followed Pedagogio's example and collected evidence of student learning. The teachers found that their new approach achieved greater success and even less failure than the lesson. Spurred on by the results, they again set to work to use the information they had gathered to improve their instruction. Thus, the cycle continued throughout Time.

No one has seen Pedagogio since ancient times, but, to this day, Man is continually reminded of Pedagogio's presence. Man still finds that he has not perfected his ability to teach, and that, to improve, he must study his successes and failures. The lesson that educators can learn from Pedagogio is that Man will never learn to be a perfect teacher, but that, even with his limited abilities, Man can learn to perfect his teaching.

* * * * * * *

"Every day and in every way I am getting better and better." That's what people said and tried to do when following Emile Coué's course for personal improvement. Yet most instructional project directors, school administrators, teachers, and textbook publishers could not repeat Coué's liturgy with any sense of honesty. Neither major improvements in teaching and learning, nor slow and steady progress are perceptible in schools today. At best, school administrators, teachers and instructional product developers would have to admit: "Every day and in every way we are barely maintaining our status quo."

Most simply stated, the field of education is stuck in a rut. Well-meaning and well-publicized attempts to introduce technology into the classroom are rare and do not begin to fulfill technologists' promises of wide-spread improvement.

Nothing seems to help. Even new and systematic approaches make very little difference in the improvement of teaching. Free schools, open schools, intuitive and humanistic approaches, performance contracting -- all have little impact. Most teachers still teach using the same basic principles and methods as during the turn of the century. In higher education, methods are not much different than those used in the Roman Empire.

To produce major changes in the field of education, instructional developers must create and perfect large-scale instructional projects. And the best way to perfect an instructional project is to employ the process of constructive evaluation.

> Constructive evaluation is a systematic process of collecting and using information to improve a developing instructional project.

Thus, constructive evaluation is characterized by its purpose, scope, and time of use. Its purpose is to improve instruction, its scope is one particular system, and its time of use is during the project's developmental stages.

* * * * * * *

An Interview

MAN: Hello, my name is John Johnson, your man on the street. And what do you do sir?

CONSTRUCTIVE EVALUATOR: I'm a constructive evaluator.

MAN: Oh, I suppose you find out how well constructives work, but what are constructives, sir?

C.E.: No, you've got the wrong idea. I help people improve their teaching projects by collecting information, by investigating.

MAN: Ah, I see, you find incriminating evidence, then tell the educational project director that you will release the evidence to the local newspaper unless he improves his teaching. Right?

C.E.: I help by showing him how to find out what students learn, and then how to use that information to improve methods to achieve better student learning.

MAN: I see. An educational project director does his job, he checks his techniques by testing student learning; then he uses that information to improve the results. That's constructive evaluation.

C.E.: Good for you! Now let's talk about improving your interviewing.

* * * * * * *

Constructive evaluation should be integrated into the development of any large-scale instructional project. Examples of large-scale projects are an industry-wide training program to teach employees first-aid; a nationally broadcast television program to teach slow readers to read; a course to teach agricultural economics by slides, tapes, and programmed laboratories; a nationally distributed course to teach parents how to use toys to stimulate a child's mental development; or a set of carefully sequenced booklets, teacher materials, posters, and games to teach reading systematically from kindergarten through fourth grade.

The field of education is plagued by many problems which can only be solved by efforts to create and perfect large-scale projects. Constructive evaluation is a systematic scheme to collect and use information to improve instruction. Testing and revising these major projects on the basis of empirical evidence will allow educators to make great strides toward more effective, efficient, and acceptable teaching.

CHAPTER II

The Big Ball of Wax: An Overview

Could Daedelus have avoided plunging to his death when trying to fly with waxed wings? Could Ford Company have avoided producing an Edsel? Could instructional designers have avoided producing their waxed wings and Edsels? Yes, they could have if they knew how to systematically employ the complete process of constructive evaluation. They had to apply the big ball of wax.

The complete process of constructive evaluation can be divided into three major tasks: 1) finding out if information is necessary to improve, 2) collecting information, and 3) using information to diagnose program strengths and weaknesses and remedy them. Thus, if you were the person responsible for a constructive evaluation, it would be your task to find program faults, determine their nature, infer their likely cause, institute changes, and check for resulting improvements. You might ask, "As the program now stands, which objectives will the students achieve and which will they fail to achieve? What unwanted, or unforeseen, results might appear? Should I revise the program? Should I eliminate, change, add, or resequence? What instructional options should I choose to remedy a fault? Which of the examples will communicate better? Which format will hold attention?"

Now, to gain a broad but meaningful view of the nature of constructive evaluation, read the extraordinary story of how David Markle, an instructional developer, created a basic first aid course for Bell Telephone. (1)

CASE

An Exemplary Course

> Markle's first aid course created for Bell Telephone employees began with six film vignettes illustrating accidents. At the most appropriate moment during each introductory film a question ("What's the first thing you do to stop bleeding?") flashed on. Next, a film explained the course procedures, which included filmed demonstrations of first aid skills. (how to stop bleeding, how to tie bandages), practice sessions, and workbook study. The course consisted of 20 films, 17 practice sessions, and 13 workbooks. Students used the workbooks to test themselves, to review material learned, to learn details, and to learn first aid knowledge which required no new skills.
>
> The entire course took one working day. The previous first aid course used at Bell took ten hours in contrast to the eight hour course created by Markle. On a wide-range test of first aid ability, untrained subjects scored an average of 26%; those trained by the previous first aid course scored an average of 47.5%, and people trained in the new course had an average score of 82.8%.

How did he do it? How did Markle create a course which was more effective and more efficient than the traditional one? He used constructive evaluation. Let's review how you could do what he did, from the beginning of development to the creation of the final product.

Starting the process of constructive evaluation is relatively easy: you can begin whenever someone gets an idea for an instructional project. You may begin by considering a problem, ("Kids are not learning to read") by stating a need, ("We need better doctors") or by proposing an instructional method ("We intend to teach reading via T.V.").

But before you start, be warned: constructive evaluation is not for every instructional project. To decide if you should use the process of evaluation, consider these points: You need constructive evaluation if 1) you are not sure that the project you propose can get your students to learn, and 2) if you want to improve your teaching methods as you develop them. You could, after all, take your chances instead, and test the product or method after it is fully developed, or you could decide not to test your method or product at all.

It is best if you begin to plan constructive evaluation at the beginning of a project. To begin, you should have some specific ideas about the method or product: a goal, a statement of the problem, or plans for a lesson. You should have the time and money to test and revise, and you should be committed to accepting and using information.

CASE

Determining the Need

> Markle decided to use constructive evaluation in his approach to instructional development because it was needed. It was necessary to find out if a first aid course could be made shorter and still teach more. Markle considered constructive evaluation to be the best way to obtain the information because he wanted to find out how to improve the new course while it was still under development. He had no intention of waiting until the course was fully developed only to discover his efforts were for naught.
>
> Markle knew he was ready to pursue the evaluation when he decided what it was he wanted to teach (an analysis of 50,000 accidents had revealed the skills which were necessary), when resources (time and money) were allocated by Bell to Markle to test and revise the developing program, and when he was given authority to use the information collected to make any revisions necessary.

First, you decide if constructive evaluation is the appropriate procedure for your project, and then you decide if you have the resources and commitment to be sure that the evaluation is likely to be completed. Next you ask evaluation questions which, when answered, will give you the information you need. You may ask, for example, "Will urban planners learn to solve ecological problems by watching films during a lecture period?" Your evaluation questions become the focal point for all subsequent decisions; you must therefore, form and analyze the questions carefully.

You should be reasonably sure that your questions are answerable within the limits of your available time, money, and staff, and you should be as sure as you can that the questions make sense. Sometimes you can detect gross inconsistencies among the parts of a question by studying the theoretical relation of those parts. It may be theoretically inconsistent, for example, to create ecological problem-solvers by requiring people to watch films. No one is going to learn to solve problems by watching a film.

CASE

Defining Evaluation Questions

The constructive evaluation question posed by Markle during the development was as follows: "How can I improve a course in first aid which will teach Bell employees basic first aid skills and knowledge at each office location in only seven and one half hours?" This question was too vague to be used to derive ways to test, choose a sample of students, and select a place to be used for a test of the course. Therefore, Markle defined each part of the question.

Once questions are stated, you should define and organize the elements of instruction described in the evaluation questions. Results may be defined as effective ("Does the audience learn?") or efficient ("Is the learning worth the cost?") or acceptable ("Does the audience like it?"). Methods and materials may be defined by their features ("This will be a readable, credible text") and by their processes ("We will present definitions, then examples, then written cases to analyze"). An audience may be defined by its members' status ("3-year olds") their traits ("highly anxious") or their knowledge and skills ("The kids have a 40-word sight-reading vocabulary"). An instructional setting may be defined by its features ("We are talking about a typical 40-seat classroom with two blackboards") or by its transactions ("There are likely to be four groups proceeding at once").

CASE

Defining Tryout Elements

Markle wanted Bell employees to learn first aid, but to evaluate their learning, he would have to carefully define the desirable results in terms of student performance. In the development of Markle's first aid course, the results were defined and organized by three methods. First, from the content of the first aid manual, Markle derived questions which students of first aid should be able to answer. Then, to find out if all important questions were stated, a grid was prepared with first aid topics on one axis (e.g. care for wounds, heart attack, artificial respiration) and five types of procedures used when giving first aid on the other (1. skills, 2. determine the action to take, 3. identify injury, 4. infering what's wrong and 5. preventing accidents). When Markle compared each topic with each first aid procedure he found that certain combinations were not covered in the questions drawn from the manual, so he added more questions which students of first aid should be able to answer.

To double check and define his requirements further, Markle drew diagrams to symbolize the steps and decisions necessary to solve some first aid problems. When Markle compared his first aid diagrams and the questions drawn from the first aid manual, he found that several steps and decisions were not covered, so even more questions were added.

The questions were placed in five categories of procedures usually used when solving first aid problems. In each category the questions were arranged according to their order of occurrence in the first aid problem-solving process. Thus, Markle had defined his results.

Markle questioned untrained potential students in order to find out what his audience already knew. Markle eliminated the questions that were answered correctly by all untrained students. Based on this procedure, Markle could reasonably estimate what his audience already knew about first aid and what they still needed to know; he had defined his audience.

The instructional setting was fairly certain -- any location housing Bell employees. But the instructional method was purposely left undefined except for one characteristic -- leanness. Because of time restrictions, the course would have to contain only the minimum amount of knowledge needed to answer the first aid questions.

When you have defined the instructional elements included in your evaluation question, you are ready to plan a tryout -- a test of the project. You will need all types of tests; a sample chapter, section or unit of your instructional methods; a sample of people representing your target audience; and a place for the tryout as much like the one in which the project will be used as possible.

CASE

Planning a Tryout

Markle had to consider results, audience, setting, and instructional method to prepare for a tryout of the new first aid course. He was ready to assess the course results by a test composed of the questions to be answered about first aid. He refined the list of questions by having potential students read and answer them. At first the program was to be tested in a laboratory environment -- a setting only vaguely similar to the place where the course would eventually be given. Later, for a more realistic test, a field environment was chosen: a place just like the one in which the course would someday be presented. The sample of people for the tryout was selected from representative Bell staff. At first Markle worked with lone individuals only in the laboratory; later, he took small groups through the field tryouts. In the earliest tryouts Markle used only questions to represent the course procedures; as development progressed, material was added until students were given a complete course presentation in the final series of tryouts.

To get the most out of a tryout you should combine your test, instructional materials, sample audience, and setting into a tryout design. Then you should conduct the tryout to collect the information you need.

CASE

Conducting a Tryout

Once Markle's tests, audience, setting, and materials were ready, he assembled them into a tryout design, and then he conducted the tryout.

There were many tryouts and, thus, many opportunities for Markle to organize and analyze the data collected, to find strengths and weaknesses of the presentation, and to generate better course materials. During the earliest tryouts, individuals were asked to respond only to questions, for the sole purpose of seeing what a student could learn from questions alone. In subsequent tryouts, individuals were given questions to answer and were given answer keys to check the correctness of their answers. When Markle observed that consistent errors, and continual requests for explanations were associated with certain questions, he added information to the program in various forms: film, text, and practice exercises.

Much later, when a reasonable facsimile of a course was available (a few black and white films, one or two practice sessions, and some workbooks), a small group of people was asked to study first aid in a typical Bell setting. Markle briefed the person who was to coordinate the materials in the program as it was used then. Markle collected data by observing the program in progress, by analyzing student's responses in workbooks, by observing students practice the skills they saw demonstrated in films, and by studying students' answers to final exam questions.

Markle tried three versions of the program in this fashion, each one more complete, more effective, and more efficient than the last. The most effective version had an instructor's manual, more films (in color), more practice sessions, and more workbooks.

Once the information is collected, you score it, summarize it, and display it.

CASE

Scoring Data

Markle scored, summarized, and displayed the data he collected. He computed 1) the time it took students to respond to test items (compared to normal reading rate), 2) the errors made, 3) the amount of time to administer the program, 4) the average score correct on the test, and 5) the deviation of scores from the average. He compared each of the results to the results of the standard first aid course. Subjective comments made by students were not quantified because they were helpful in the form in which they were given.

Once the data were organized, a number of things happened in quick succession. Markle identified the strengths and weaknesses of the program, he hypothesized which instructional factors were contributing to the course's strengths and weaknesses, made revisions for the course, made priorities among modifications, and finally revised the course.

Once you have organized your data, you compare your results to some desired achievement level to identify the strengths and the weaknesses of the course. Then you make hypotheses about what you believe contributed to the results. First you might hypothesize that certain examples and exercises in the course might be affecting the results. You might hypothesize for example, that the ecology films did not provide sufficient practice on air pollution problems. Several hypotheses like this are likely to be formed, but you probably will not be able to act on all of them. You will have to rank them in order of importance.

When your hypotheses are backed by strong evidence, you are likely to generalize about the relationships of aspects of your course and certain results. Often these generalizations become operating rules which project directors use to form or to revise their courses. An operating rule might be this: "When introducing basic concepts use an example with salient attributes."

If you decide that modifications are necessary to increase the strengths and reduce the course's weaknesses, you do so. After you have made changes in the program, you will have to decide if you wish to test again either by constructive evaluation, or by some other approach. Or you may simply decide to use the course as it is.

CASE

Finding and Explaining Strengths and Weaknesses and Revising the Program

According to a cut off point set by Markle, too many students made consistent errors in answering certain questions on early drafts of the materials: they said, incorrectly, that frostbite should be rubbed, that feet should be elevated in a case of head injury, that an injured person should be removed immediately from the wreckage of a car. Markle hypothesized that the reason for errors was either insufficient information or ambiguous course content, and he decided that he should add to and clarify the content.

Markle found other course weaknesses. Students, for example, read some workbook segments at a rate slower than the average reading rate. Markle's hypothesis was that the writing in the workbook was ambiguous; so his changes consisted of clarifications.

The first tryout took students twelve hours to complete. Markle hypothesized that slow student progress was caused by too much redundancy. To cut redundant content, Markle removed presentation and practice sessions for those first aid questions that were answered correctly early in the program. If a

> question in the program was consistently answered correctly, to save time, Markle converted the question to a statement.
>
> Markle discovered, through observation, that the person who administered the course often became confused. Markle thought the reason for his confusion was the overwhelming number of documents he had to use. As a result, in a revised version all instructions were included in one booklet.
>
> During the practice sessions and on the final test, students could not imitate the skills demonstrated in the film. Three explanations for this were offered: that there may have been visual or audio distraction and interference in the film presentation, that in each film too much material may have been presented at one time, and that the instructions for skill practice may not have been appropriate preparation for the test.
>
> Markle made the following changes. First, when an important visual was to be studied, the narrator said nothing; when an important statement was to be made, the visual was kept still or darkened. Longer films were cut into segments and more practice sessions were given. Finally, the phrasing of questions for skill practice was made the same as the phrasing of test questions: for example, instead of saying "Apply direct pressure to the wound" for practice, the format read "Do the first thing for bleeding" just as it did on the test.
>
> After each set of changes were made, Markle had to decide if he was ready to release the program for use or if another test cycle was appropriate. Ultimately, Markle ran through at least seven distinct cycles.

Thus, the test cycle may begin again and you may determine the need for constructive evaluation, then you may ask questions, define instructional elements in the question, choose tests, select samples of the audience, pick instructional segments and arrange for a test setting, design and conduct a tryout, organize the data, find strengths and weaknesses, and make changes.

> In this case, Markle demonstrated the best use of constructive evaluation. Step by step he used constructive evaluation techniques to collect and use data to improve his developing course of instruction. Not only did he produce a highly effective system, but he also increased the efficiency of the course.

Although Markle's study is a real case of constructive evaluation on a large-scale project, you might get the impression that the process moves along in an orderly fashion. Sometimes describing a process systematically has this effect.

Often statements describing dynamic processes like constructive evaluation are deceiving because they leave out the underlying principles which give such a process its vitality. There are four principles which provide the working basis for constructive evaluation.

Constructive Evaluation Requires Commitment

To continue the process, constructive evaluation re-requires a project director's commitment to use the information collected.

The process can continue only if a project director wants to know more about the effects of his work, is committed to use information collected, has the time, money and staff to gather data, and then puts the data to use. Commitment is essential to prevent the constructive evaluation effort from being wasted.

Constructive Evaluation Requires Continuous Reporting

Before, during, or after any part of the process, a progress report may be in order.

The function of constructive evaluation is to get information -- direct and complete answers about the effectiveness of a course -- to a project director when he needs it. Reporting is most important during the early phases of development when project directors are usually able to make changes rapidly and easily.

The Constructive Evaluation Process Requires Its Own Analysis

A project director should continuously study the constructive evaluation process as it flows, to find ways to improve it.

A director should analyze the process of constructive evaluation; its testing procedures, organization, analysis methods, and reporting techniques. He may suggest revisions for any part of the process.

Constructive Evaluation Involves Complex Interaction

The steps of constructive evaluation interact with each other and with other entities and processes.

Interaction is the give and take between two entities. Person A acts; his friend B reacts; person A reacts to B's reaction. As you conduct a constructive evaluation you act and react to many events, processes, and people.

You will have to react to new circumstances. Each successive use of constructive evaluation is a different case. The procedures in the second cycle may differ from the first cycle because you will be testing a new version of the course, and because you may be using perfected testing procedures.

You will have to react to several aspects of constructive evaluation at once. For example, while a polished draft of a workbook is being written, you may be asking subject matter experts to look at, and pass judgments on, an existing rough draft of the workbook; in addition, you may be asking one staff member to analyze data gathered through interviews of students who read the existing workbook draft, and you may be asking another staff member to plan the collection of data for the polished workbook draft.

You will have to react to production deadlines. When possible, you will have to schedule tryouts so they fit to your production calendar. To be ready for a tryout, for example, when a functional version of a course is available, you must have a test chosen, a sample audience selected, a place for tryout set, and a method to organize and analyze the results planned.

You will have to react to your staff's performance. Your staff's abilities will influence their effectiveness in using an information-gathering technique. You may be likely to change your plans even when the plans include a procedure tried often in other evaluations. Even though an interview technique may be well researched, for example, you may find your staff's interview skills insufficient to use the technique.

You will have to interact with the institution in which you work because constructive evaluation is affected by the institution in which it is taking place. Changing the production schedule of the instructional television show, "The Electric Company," exemplifies the interdependence of parts of an instructional development system. For financial reasons, the taping schedule was changed from a full year of production to two separate three-month periods, one during the summer months. This schedule change forced by the institution, could limit the amount of available information about the effects of a previous season because the schedule change could mean planning for a new season before the old one was done. In addition, the summer taping could limit the number of available school centers in which new segments could be tested.

As part of many complex interactions, you will have to take into account the instructional development process. Although techniques vary, a project director begins by gathering data on the need for a course, data on student background, and data in the classroom; then he derives objectives, and last, he creates methods and materials. You will have to be aware of the stages of development of the course to construct your evaluation questions. For example, you will have to know the course objectives to know what results to measure, and you will have to study the materials to know which portions to select to test.

* * * * *

Constructive evaluation is a systematic process which you can learn to use to improve your instructional projects. But each part of the process must be mastered for the whole system to work well.

Had Daedelus (our unfortunate hero who died by coming to close to the sun with his waxed wings) known about constructive evaluation, he might never have flown before he had a sound method of flight. I wonder how many project directors are stepping off cliffs now without any idea if their projects will fly.

The Process of Constructive Evaluation in Brief:

--You determine the need for using constructive evaluation to collect information and you decide if you are able to finish the process.

--You form evaluation questions which, when answered, will provide the information needed.

--You choose, define, and organize the results, audience, method, and setting derived from the evaluation questions.

--You form a test of the instructional method or product. You choose a sample of the course, the audience, and the setting.

--You combine the instructional elements to create a tryout design.

--According to the tryout design, you conduct a tryout to collect the information needed.

--You organize the data.

--You identify strengths and weaknesses of the course.

--You hypothesize which factors contribute to the acceptable and the substandard results.

--You extract operating rules from the preceding analysis.

--You generate modifications of the course.

--You make priorities among modifications.

--You modify the method or product to the extent of existing resources.

--You recycle or do a final evaluation.

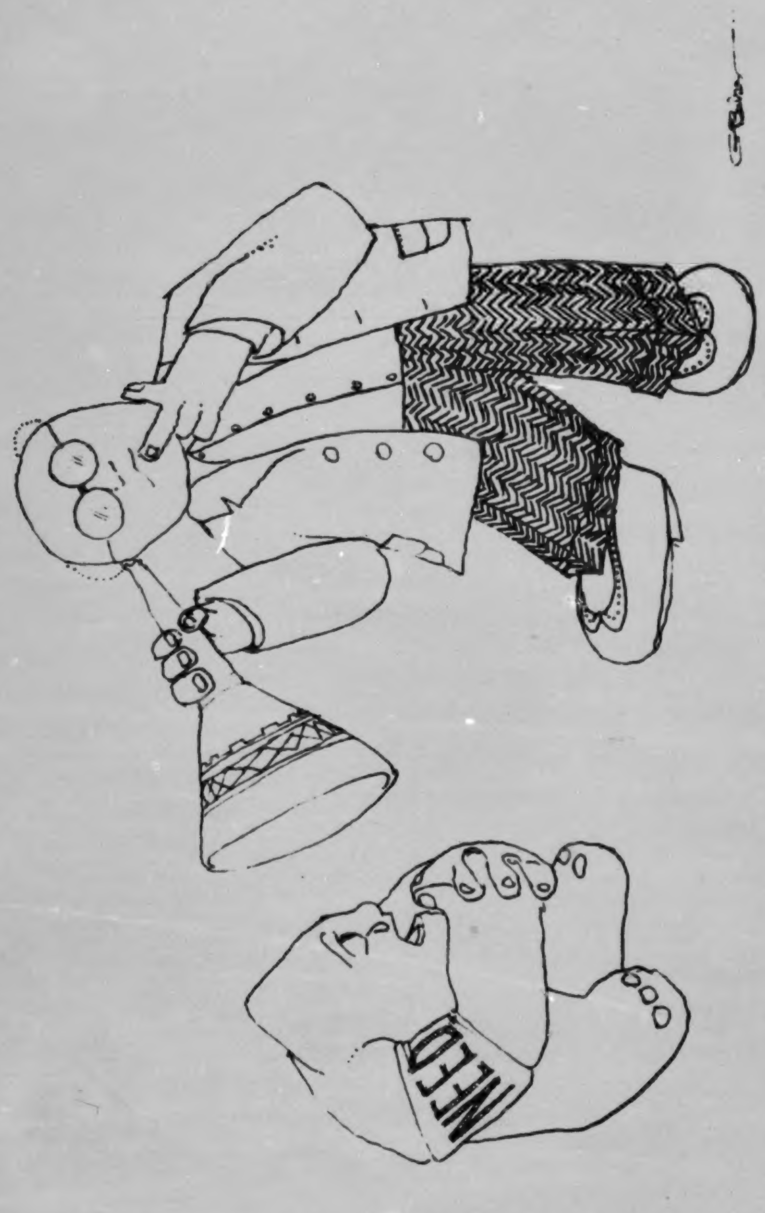
The Principles of Constructive Evaluation:

--Constructive evaluation requires commitment.

--Constructive evaluation requires continuous reporting.

--The constructive evaluation process requires its own analysis.

--Constructive evaluation involves complex interaction.


NEED

CHAPTER III

Mental Gymnastics: Determining the Need for Constructive Evaluation

As the director of a large scale instructional project, you will have to be a mental gymnast to manage all the problems which are likely to be on your mind. To do so systematically, you will have to do mental somersaults, twirls, and jumps. You will have to juggle facts and theories; you will have to leap from one decision to another, and you will have to shoulder heavy plans while carefully balancing your time and money.

To perform gracefully and successfully, your moves must be both necessary and carefully planned. Before performing a major and complicated mental routine like a constructive evaluation, you must be sure that the routine is essential and that you are fully prepared to complete it.

Three Questions For Two Decisions

You will not want to conduct constructive evaluation for every project you produce: it is not always necessary. For each project, you will have to decide, first, if you need the kind of information a constructive evaluation can supply and, second, if you are fully prepared to conduct the evaluation. To make these decisions for your project, ask yourself three questions. First ask, "Do I need information that will guide my work while creating this new teaching method or set of materials?" Second, ask, "Should I use constructive evaluation or should I use some other process to secure the information I need?"

Third, ask, "Is it possible to conduct a constructive evaluation? Should I plan to begin and carry out a constructive evaluation?"

* * * * *

Do I Need Information That Will Guide My Work While Creating This New Teaching Method?

The evaluation of instruction is an undertaking far too difficult and complex to be handled by intuition alone. Subjective judgments of the worth of an instructional system have been consistently invalid and unreliable (1). (2, 3) For example, the effectiveness of seven versions of an instructional program was tested by asking students to learn from each one and then take an achievement test. Twelve teachers trained in a course on programmed techniques were told to read and rank order the materials according to their predicted effectiveness. Their prediction correlated - .75 with the empirically derived student scores; in other words, their predictions were the opposite of the results found. (Other instructional developers have reported similar events.) (4, 5, 6, 7)

Most educators rightfully assume that information provided by evaluation will not take the place of creative intuition, but not all educators are aware that creative intuition does not take the place of information provided by evaluation. Certainly intuitive insights have contributed to many interesting and important creations, but apparently they contribute little toward judging the worth of a project. Decisions that involve student learning should have a more concrete basis than a person's intuition.

If you have doubts that your project will result in learning, then you must collect some information to increase the probability that your project development will be successful.

The primary criterion for you to use to determine your need for information for a given project is your doubt about the probabilities that students will learn from the instruction.

The following case of a large scale project illustrates the mental processes that you can use to prepare for a constructive evaluation. I will show how the producers of the Health Show at Children's Television Workshop answered the questions and made the decisions preparatory to the evaluation of their project.

CASE

Doubt About Learning

During 1972 and 1973 at the Children's Television Workshop a new instructional television show was being planned. The development of the show began with the idea that adults needed to know how to take better care of their health. Early in their planning, the producers realized they would need constructive evaluation, and were ready to begin that evaluation process. Let us consider how they arrived at that conclusion. (8)

The health show staff members had some doubts about their ability to teach things dealing with health. Health show producers and evaluators did not know the answers to some major questions:

1. Will the material interest the audience?
2. Will the material be understandable?
3. Will it be remembered?
4. Will it be credible?
5. Will it lead the viewer to take appropriate action?

The producers had to collect information to answer these questions.

If you have doubts about the probable success of a project, and you believe the learning in question is important, then you must collect information to guide your efforts.

How important is learning a particular attitude, skill, or concept? That depends on what is needed in the community, the school, or the classroom. A need is a perceived discrepancy between what someone wants and what someone has, what someone knows and what someone does not know, what could be and what is likely to be. To specify a need you must find out what teachers and community people expect. (9, 10)

You must be sure to find the real needs, not merely the expressed ones. (11) For example, a teacher once expressed a need for more laboratory space because he felt he did not have enough space for his students. After analysis, the apparent lack of space turned out to be due to other factors: the teacher was spending lab time to lecture to small groups of students on how to use the equipment. His real need was to find an efficient way to use lab time and thereby save space. By writing out precise instructions on how to use equipment, more than 25% of the lab time was saved, and he had more laboratory space than he could fill.

Force yourself to rank order the statements of need, not according to the size of the perceived discrepancy between what is and what should be, but according to the importance of the difference. (12) The importance of the discrepancy will determine the size and extensiveness of your project evaluation. For example, if a controversial and great change is likely to be produced in the lives of many individuals by your instructional project, then it deserves rigorous and extensive evaluation.

CASE

Importance of the Project

At C.T.W. the process of developing the health show started by finding out there was a need for such a television series. The producers and evaluators interviewed one hundred and seventy health experts. These interviews convinced the producers that their potential audience did not know such basic information as normal body temperature, that medical standards in our country are low, and that there are inequities in medical care. Based on this information, researchers wrote a health show prospectus which stated:

> "Despite his need for health information, the average layman suffers from a profound lack of accurate knowledge about even the most elementary principles of good health. He also has many misconceptions about health and health care. A 1970 Lou Harris survey for Blue Cross showed that over half the public wants more health and medical information."
>
> "...In 1969 our infant death rate exceeded that of 14 other countries. Non-white American babies die at a rate nearly double that of white American babies. ...American mothers die in childbirth at a rate exceeding that of 11 other countries. The death rate for non-white American males between the ages of 40 and 50 is double that of white American males. ...American males have a shorter life expectancy than the males of 19 other industrial countries. American women have a shorter life expectancy than women in 16 other industrial countries." (13)

The health show staff at C.T.W. concluded from their investigation that they were going to teach something important and, because of its importance, they also concluded that they must collect information to insure that the show would develop to its fullest potential.

* * * * *

Should I use constructive evaluation or should I use some other process to secure the information I need?

You may decide to produce a project and evaluate it only when it is in final form, or you may decide not to evaluate it at all. But if you desire information that will lead to the improvement of the project, and wish to collect and use that information before the project is fully developed, then you have decided to engage in constructive evaluation.

> If you wish to collect information for improvement during the development of the project, constructive evaluation is the process to use.

To find out if constructive evaluation is the appropriate process for securing the information you need, you can compare your needs to the uses and functions of constructive evaluation. The constructive evaluation process provides reasons for revising an instructional method or product; the constructive evaluation process often contributes to instructional theory; and the constructive evaluation process saves time and money.

Let's consider each function in detail.

> If you want to make improvements on your project based on sound reasons, then collect your information by constructive evaluation.

While it does not provide pat formulae, the constructive evaluation process does provide reasons for making revisions which are likely to improve a project. And every project requires revision because instructional methods and products are complex and unpredictable. No amount of subject-matter knowledge or technological wizardry can counteract all the errors likely to be present in an instructional program. (14) A notable example of this principle was the instructional film Freedom and You, which had an effect opposite that desired by the producer. Those who viewed the film became less politically interested, (15) while the producer's intent had been to promote the opposite attitudes. Freedom and You is not an isolated case; every teacher can recall some well-intended experiment that boomeranged.

Although many project directors believe that one teaching method is better than another, there is little convincing empirical evidence that this is true.(16, 17) At this time it is also difficult for a project director to generalize about the effects of a given method from one educational setting to another. Each educational institution's situation -- its goals, population, teachers, materials, methods, and community support -- is likely to be different. Thus, it is entirely probable that a project may be relatively more effective in one school than in another; the differences being the result of differences in situation, not educational technique.

If you wish to contribute to instructional theory, then use constructive evaluation.

Even though the results of constructive evaluation are specific to the single program you are testing and may not be generally applicable, they may still contribute to instructional principles. You may discover the attributes of an effective program, attributes that can be featured again and again (meaningfulness, activity, humor, suspense, saliency).

In addition, you may be able to identify relationships between types of individuals and forms of instruction -- which individuals benefit and which do not.

The results of a constructive evaluation may lead you to make hypotheses for experimental research. From your data you may suspect the effectiveness of a particular variable -- the structure of a message, for example; but a controlled experiment has to be done to be sure of the variable's effect.

> If you do not wish to risk a total loss of project expense, but wish instead to save money in the long run, use constructive evaluation.

Education, good or bad, is not cheap. If a large-scale instructional project is important enough to justify expense and effort, it certainly deserves testing. The benefits of constructive evaluation are worth the costs of testing because the process may prevent the production of a course which turns out to be a total loss. Constructive evaluation increases the likelihood of producing more effective and, possibly, more acceptable courses, although these may be fewer in number.

CASE

Choosing Constructive Evaluation

The health show staff adopted the technique of constructive evaluation as the process that fit the producers' information-gathering needs: they needed a way of improving the pilot show during the eight months it was being developed. They wanted a rational and systematic way to make production decisions; they wanted to learn how to influence health-related behaviors via television; and they did not

> want to develop a program only to discover, once it was broadcast, that many aspects of the project were wastes of time and money. (18)

In summary, if you desire a rational basis for improving a developing instructional method, if you want a process which may yield principles which aid other endeavors, and if you wish to save time and money, then your choice of an information-collection process should be constructive evaluation. But your choice of process does not end your mental gymnastics; you still must decide if you are prepared to begin a constructive evaluation.

* * * * *

Should Constructive Evaluation Plans Begin?

You may be able to begin a constructive evaluation when your instructional ideas are clear, your resources are sufficient, and your attitudes toward change are positive.

> If you have specific ideas for instruction, you will be able to start asking sensible evaluation questions.

To plan an evaluation you must have some idea of what you want to test. You should know enough about your student population, the specific skills and knowledge you want them to learn, the materials you intend to use, and the instructional setting in which it will be used, to be able to ask specific questions about each one. The more specific the ideas, and the more varied their expression, the better.

Goals help you form evaluation questions about the effectiveness, efficiency and acceptability of the project.

Goals are expressions of your instructional intent; they represent your desired results. A goal is a statement describing the change you require in a student; it can be general, or it can be quite specific and observable. You can say, "A student will understand health principles"; or you may be more specific and say, "In a mock emergency health situation, such as child poisoning, a student will apply procedures to resolve the problem so that the child's state of health is stabilized or maximized according to the Guide for Home Health Emergencies."

There are many uses for goals in constructive evaluation. If you work with a team to produce a project, the producers, technicians, and evaluators can coordinate and direct their activities more accurately by keeping a goal in mind.

Goals are the basis for most of your decisions. Goals enter into the formation of evaluation questions such as "What health behaviors do I want learned?" Goals also help you determine measures: "I will need tests to know that health goals have been accomplished." In addition, goals help you in the organization and analysis of data: "I wonder which goals were reached and which were not?"

You should be aware of some tricks of the trade which relate goals and constructive evaluation:

It is all right if goals are vague at first; you can make them more precise later. The health show goals, for example, were first stated very broadly and later became more specific.

If you are producing a project with a team, do not press for goals too early; you may kill creativity and initiative. Check the extent of the team's progress first. If they have some general notion of purpose, then you can ask for goal statements specifying what students will be able to do.

A list of goals shows what is worth special thought but not exclusive attention. (19)

Make priorities among goals. The priorities will show which goals should be pursued and to which ones resources should be allocated. (20) Priorities also determine effort, but priorities do not guarantee success. (21)

Long term goals, such as "changing health attitudes," are usually not testable during constructive evaluation; there is usually not enough time, for example, to obtain and respond to evidence of attitude change (which may take weeks or months). But, if you have time, a particular behavior, such as getting a V.D. check-up, might indicate some change in the direction of a long term goal. If you observed a series of similar behaviors over time, you may be convinced that an attitude had changed.

CASE

Stating Goals

To meet the community need for knowledge on health, an entertaining series of 26 one-hour shows was proposed by C.T.W. It was planned that these programs would give the American public accurate, useful health information, would show that there are actions the average citizen can use to take better care of himself, would make the health-care system better, and would teach specific behaviors, such as getting immunizations or changing eating habits. Specifically, the major goals of the C.T.W. Health Series were to:

> "1. Instill in the viewer a greater concern for his personal health and to encourage him to become aware of his responsibility for maintaining his physical and mental well-being.
> 2. Give the viewer useful and reliable information for both good health care and prevention of illness.
> 3. Help the viewer effectively use the health care system for the benefit of himself and his family." (22)

To decide what specific information, attitudes, and behaviors should be taught, task force meetings were held. The Children's Television Workshop invited subject-matter experts, community leaders, and members of potential target audience to a series of discussions to ferret out the most important and most needed content. A tentative list of goal topics was made.

The task forces were drawn from the general goal areas suggested by the first set of interviews. (See list) The task forces considered pre-natal, infant, and child care, adolescence, modification of habits, chronic disease, family planning, access to health care services, and death. From these topic areas, rough goals were extracted according to certain criteria:

> "1. The importance and significance of the subject as defined by its prevalence and the force of its impact on the function of the individual or his family.
> 2. The degree of public interest in the subject.
> 3. The extent to which an individual can do something about the problem himself.
> 4. The extent to which a doctor or other health specialist can do something about the problem.
> 5. The potential for effective television treatment.
> 6. The susceptibility to measurement of the impact of a program on the viewer's knowledge, attitudes, and actions."

General Areas suggested by initial interviews

I. Basic Prevention and Health Maintenance

Self-abuse
- smoking
- alcohol
- drugs and pills

Parenting:
- prenatal care
- immunization and childhood diseases

The Human Body and Its Maintenance:
  general nutrition
  accident prevention
  sex education and venereal disease
  vision
  home care of the bed patient
  exercise
  physical and dental examinations
Chronic Diseases:
  arthritis
  allergies

well-baby and child care
dental care
family planning
accident prevention
skin and hair care
dyslexia and brain damage
family relationships
genetic counseling

Mental Health

Advances in Medical Science

Sickle Cell Anemia

Death

Ten Leading Causes of Death: how to recognize and decrease probability of occurrence.

Common complaints: appetite, insomnia, concentration, memory, constipation, fatigue, nervousness, boredom, phobias

II. Access: How to Find Appropriate Care and Make Better Use of the Health Care System

Doctor/Patient Relationship:
  When to see a doctor
  How to find a doctor
  Should doctors make house calls?

Responsibility of the patient: What should patients reasonably expect from doctors?

The User as Part of the System: participation and influence

New Forms of Care:
  Group practice
  Neighborhood care: community health centers; their relationship to hospitals

  Allied health personnel

Emergency Medical Services

III. Community and Environmental Problems

Pollution: air, water, and noise
Lead Poisoning
Rats

IV. Paying for Care

Why are costs so high?
Prepayment, Medicare and Medicaid

Health Insurance: why and how to buy National Health Insurance: care as a right.

Each task force approached its concern in its own way. During the adolescence task force, the following questions were asked: What are the concerns of adolescents? What information do they have and what do they need? How can they be reached and motivated? The task force members found that adolescents are primarily concerned with their self-concept, their peers, and their own feelings of normality, and that physicians rarely discuss these matters with adolescents. Some of the tentative goals suggested were these: that parents know the common concerns of adolescents; that they be made more willing to communicate with their adolescents, and that they become aware of the role imitation plays in the relationship of their own drinking and drug-using behaviors to adolescent alcohol and drug abuse.

The task force members on child care suggested the following tentative goal areas: after completing the program, viewers should be able to prevent accidents, should be able to label their own feelings and attitudes toward parenthood, should know the normal range of child behaviors, should know the development of language and early learning, and should know how to provide a well-balanced diet.

The members of the task force on access to health care services promoted the following goals: the audience should know their rights as patients, and should know how to react in an emergency or a crisis.

The specialists in modification of habits considered the topics smoking, drinking, overeating, and drug use. Goals on drug use were reserved for a later time because of the oft-noted boomerang or reverse effect of drug education programs and because very little knowledge about how to change drug use habits is presently available. Because of unsuccessful previous attempts to change unmotivated people, the audience chosen for habit change was composed of those people already on their way toward change. The ultimate goal of the specialists in habit modification was to support the change, to help him continue to change. Other goals included knowledge of alcoholism, its signs, and its causes.

The goals relating to obesity overlapped with nutrition, but the emphasis was on how to deal with stress and crisis in ways other than eating. Also task force members felt that viewers should learn how to employ physical, medical, and social alternatives to change eating habits.

Heart disease, cancer, and diabetes, and the relation of nutrition, stress, and exercise to these chronic diseases, were the focus of another set of task force goals. The viewer would learn to relate these factors to his life and act in a way which would prevent chronic disease.

Members of the family planning task force wanted the audience to learn how to choose a family size best for their own values and beliefs, and then, after making a decision, know which services were available.

> The tentative goals proposed by the task force on death focused on knowledge of the concept "death with dignity." Viewers would learn to talk to a dying person about death, would learn to allow a dying person to make decisions until he dies, would learn how to explain death to a child, and would learn what behavior to expect from a child when someone dies.

Even after all this effort the health show staff may find that other goals are more important than those already listed.

If you have stated specific course content, you will be able to ask precise questions regarding the relationship of your course's content to your course's results.

In the same way that you can use goals to direct your observations to certain behaviors during testing, you can also use concepts, principles, facts, and skills to orient your view to even more specific, more efficient questions and tests.

CASE

Selecting Course Content

> At the time this manuscript was prepared, the health show staff had organized some of its content according to goal areas. But they could have specified and analyzed their content in a number of other ways.

One way to structure content is by charting the relationships among ideas. Any subject matter is a system of interlocking, supporting parts, and can be listed, charted, or diagrammed as in the following example. The list includes general content categories and examples of questions the health show staff might want people to master.

DEFINE OR STATE A CONCEPT OR PRINCIPLE.

(What is a well-balanced diet? What will a well-balanced diet do for you?)

RECOGNIZE AN EXAMPLE OF A CONCEPT.

(Is this a well-balanced diet?)

DISTINGUISH BETWEEN EXAMPLES AND NON-EXAMPLES OF A CONCEPT.

(Which of the following are examples of a well-balanced meal?)

GIVE EXAMPLES OF A CONCEPT.

(Give an example of a well-balanced diet.)

RECOGNIZE CORRECT AND INCORRECT APPLICATION OF A PRINCIPLE.

(If a diet is balanced, a child will grow fat; true or false?)

PREDICT CORRECT APPLICATION OF A PRINCIPLE.

(If a diet is balanced, your weight will____________.)

APPLY PRINCIPLES TO SOLVE A GIVEN PROBLEM.

(Someone points out that a health problem exists and you derive a solution from principles of health: the doctor tells you that you are overweight, and then you reduce intake.)

RECOGNIZE A PROBLEM AND SOLVE IT.

(Find that a health problem exists and derive a solution from principles: you discover you are overweight and reduce intake.)

PRODUCE A NEW, UNIQUE, OR CREATIVE SOLUTION TO A PROBLEM.

(An original solution to a health problem: you find a new way of dealing with nonorganic obesity.)

Suppose for example, a health show staff member starts to test in the area of nutrition by asking about vitamins and nutrients, then continues to test by asking about the presence of vitamins in certain foods, and then continues to test by asking about the relation of a diet of certain foods and good health. That staff member can reduce testing time by continuing to test until a student fails to answer one item correctly. He can stop testing a student then because he has pinpointed the student's abilities. A student cannot answer questions about the presence of vitamins in certain foods if he does not know what a vitamin is.

You can form another sort of content chart by placing ideas on opposing axes of a grid. Markle used a grid of first aid content areas (care for wounds, heart attack, artificial respiration) and first aid steps (determine the action to take, identify the injury, infer what is wrong, etc.) in the first aid decision-making process, to organize an extensive amount of course content. The health show staff members who are interested in emergency health care could follow Markle's lead.

The health show staff could specify many of the skills and a number of the ideas related to the goals by making lists of steps or diagrams -- essentially sets of instructions -- which show how to provide a nutritional menu, how to decide to go to a physician, and other actions. If they were to specify how to treat a dying person, for example, they might be able to identify the steps and ideas required for treating a dying person which might be lacking in a typical viewer's experience. They might look at a list of steps and ask specific questions, such as, "Do the viewers know how to deal with a dying person by talking with him about death and helping him settle his affairs?" Because of this process, their analysis of constructive evaluation data should be more complete, directed, meaningful, and useful.

With the aid of a precise description of how to plan a nutritional diet, for example, you should be able to construct a diagnostic test which measures any small and related operation, (how to keep vitamins in food), or concept (vitamins, balanced diet) needed to plan an adequate diet. A diagnostic test may help you identify those student errors which are caused by a failure to learn certain basic ideas or skills.

The descriptive and analytic techniques used to organize content force you to look at the hidden dimensions of your goals: the inner workings of the subject to be taught. When you know the ideas and skills contained in your course, you can ask specific questions about the relationship of course content and course results.

> If you can consciously apply a set of theoretical principles to your instructional projects, you can ask discriminating questions about your project's effectiveness.

The process of constructive evaluation is a cross between trial and error and application of scientifically-based theory. Trial and error is simply using some instructional procedure without any regard for why it is done the way it is, with no concern for why certain tests are used to measure the effects of instructional procedure, and with no questions about effectiveness on which to focus observation. A theory -- a set of scientifically derived principles -- can be used to determine how an instructional system is built, and what questions are asked about its effectiveness.

The problem then becomes how much trial and error to balance with how much theory in constructive evaluation. (23) You should employ enough theory to help develop your original instructional product, to form your evaluation questions, to draw inferences about strengths and weaknesses, and to generate modifications. (24) While using theory as a base, you should remain open and flexible and ask questions which contradict theory, employ tests and observational schemes which allow for unpredicted results, make an occasional wild guess on limited data, and experiment by trying to reach objectives which don't precisely fit any theoretical mold.

This confirms the notion that you should go beyond the base that theory provides and that a good deal of constructive evaluation is discovery. In fact, it emphasizes the point that the process is a compromise between the strategy of trial and error and the application of scientific knowledge.

What if no theory exists to help direct your project's development? In that case you could build your program intuitively, and later, to form a theory, try to infer why instructional decisions were made.

The important point is that what you believe about learning will determine your evaluation questions, your tests, and your revisions, even if you do not express your theory. You should detail your personal theory because it is easier to plan constructive evaluation when you know the assumptions upon which your entire instructional method is based.

CASE

Applying Theory

The health show staff made extensive plans to use available theory to construct the segments for the program. Changing health behaviors developed over a lifetime is a difficult task, and the health show staff realized that many influence strategies would have to be used: some had already proven successful; others were on an experimental basis.

The main thrust of C.T.W.'s program is to develop in people reasons to be healthy, and then use those motives to change behaviors. Psychologists have been successful in reinforcing existing behaviors and changing uncertain views into definite attitudes. But there are some gaps in theory; for example, what makes a nurse or a doctor or a cartoon character believable, what makes a message believable, how do laymen interpret health statistics, how does a viewer make the leap from belief to action, what persuasive techniques make people act differently, and how do you make a serious topic, death, for example, attractive and attention-getting?

Numerous influence strategies have been suggested. Teach people what action steps to take. (25) Create a motive by showing the negative effects of a disease; show what can be done; demonstrate how it is a personal threat to the viewer and illustrate the precise behavior to use. Ask for a token action or a token commitment from the viewer which may influence later behavior, such as saying "I will go to get a chest x-ray." Appeal to the viewer to watch on behalf of a friend and to convince the friend about what he has learned. Repeat the theme to help a message get through.

To employ many of these strategies, content, including appropriate responses and sequences of actions, must be defined by techniques detailed in the last section, such as diagrams. No single strategy should be used for the whole show; the optimum combination would be found and used.

Some other theoretical decisions about the show's makeup have already been made:

> "It will talk to the viewer in terms of his needs, his feelings, and his perceptions. It will recognize the relationship between his lifestyle and his health. Most important, it will do this in a _positive_ way, emphasizing what he _can_ do to feel better.
>
> Good health will become an integral value -- something to strive for because it helps one feel better, function more effectively, and attain greater fufillment." (26)
>
> "Though we have not made final decisions about the style and format of the series, we plan to use a broad range of television techniques to make the programs exciting: drama, comedy, satire, animation -- even short 'commercials' about good health practices." (27)

The health show staff could now ask more specific questions: Will teaching action steps result in action on the part of the viewers? They could rationalize the need for tests of characteristics like believability. When they test a pilot they might be better able to infer what instructional factors add to certain results and hypothesize what to do to improve a segment.

If you have precise instructional plans, you can ask precise evaluation questions which link instructional procedures and desired results.

To be ready to ask questions about your instructional strategies, you will need detailed, rough, instructional plans. Your instructional specifications could include a specific goal, lists of information to be given to a student, practice provided for a student, practice tests, attributes of acceptable and unacceptable student performance, and student prerequisite abilities, skills, and concepts.

Good examples of instructional plans are the writers' notebook of the Children's Television Workshop, and the writers' notebook of Bilingual Children's Television. Entries in the writers' notebook at C.T.W. which are an extension of their goals, conform to four criteria: a focus on the psychological processes in a goal behavior, use an extension of a child's own experience, promotion and creation of ideas by giving highly divergent examples, and provision of general, and sometimes specific, suggestions. The following example is an excerpt from the "Sesame Street" writers' notebook. Based on this sample, the researchers at C.T.W. could make specific questions, such as "would the sorting (number three) technique increase the likelihood of a child's being able to make a letter sound when shown a letter?"

An Excerpt from the "Sesame Street" Writers' Notebook IV

The following is a list of suggestions for teaching some of the symbolic representation goals:

LETTER SOUNDS

1. Use closeups of people's faces (not puppets) saying letter sounds. It is important for children to see the position of the lips in producing various letter sounds.

2. Play games which require the child to supply words which start with a particular letter sound.

   Ex. 1: A story or a poem is read to a group of children but certain key words are missing. The children are asked to supply the missing words and are given the clue that all the missing words begin with a given letter sound.

   Ex. 2: An alternative to the above game is to present the children with two or three pictures of objects that would be equally appropriate to fill the blank in the poem or story and ask the children to pick the one that begins with a given letter sound.

3. Sorting or classification could be done with initial letter sounds.

   Ex. 3:

   Name each picture and ask which doesn't belong -- After pointing out that truck does not begin with the S sound 'ssss', read the three 's' words again and emphasize the S sound at the beginning of the words. (28)

Bilingual Children's Television, a company formed to produce a nationally broadcast bilingual children's program, puts it instructional plans into what they call a pre-script. Each show is organized around a theme, so that the first item in the pre-script is an explanation of the theme. If the theme were "The Market Place," its appearance, location, activities, and contents would be described in detail. Historical and miscellaneous notes would be included; songs and recipes would follow. Next, instructional segments would be described according to goals arrived at by a goal grid. An example of a pre-script follows. Based on the ideas in the prescript, an evaluator could ask "How many examples of each type of exchange must be shown before a child will be able to recognize the three models of exchange when he sees them?"

Excerpt from Pre-script* -- The Market Place by Bilingual Children's Television

(29)

CURRICULUM

Goal Matrix: ____________ Ref. No SOC 4 G (1b)

Theme: ____________ Language: English ____________

Spanish ____________

LEARNING TASK

To discover that one of the ways of acquiring needed goods is through exchange, and that there are various modes of exchange.

SEGMENT DESCRIPTOR

The segment will demonstrate that exchange, in its simplest and most basic sense, has three modes:

- a) goods for goods.
- b) services for goods.
- c) money for goods.

SPECIAL INSTRUCTIONS

1. The exchange of goods for goods, and money for goods should be clearly and visually demonstrated. The actual exchanges should be made in close-up shots, where goods are shown to change hands.
2. The theme of market place should be exploited here with these exchanges taking place in open air mercados, with street vendors, in tienditas, etc.
3. The "shoppers" should carry their own baskets or shopping bags as is the custom in Latin market places: Focus on the bartering which is customary also in market places.

EXAMPLES

1. Show someone in a market place shopping from stall to stall. He pays money for goods. Exchanges should be simple, and the actual exchange should be highlighted.
2. Show two peddlers exchanging goods for goods.
3. Show a character who wants to buy something but has no money or anything to exchange. The shopkeeper tells him he will give him some of the needed commodity in exchange.

(SEGMENT 30)

Another form of a teaching plan is called Instructional Specifications (I.S.); it includes an objective, a cue (a major rule or idea), a mastery item (practice and test of the objective), limits (a criterion used to judge the adequacy of performance on the mastery item), and entry skills (prerequisite abilities required to begin the unit). An example on the basis of these specifications follows. An evaluator may be able to ask "Will a person based on a definition alone be able to discriminate among statements of observation and inference?"

"Sample I.S.: Identifying Statements of Observation

Objective: To identify statements of observation given statements of observation and inference, and the objects or events to which the statements refer.

Cue: A statement of observation tells what you see.

Mastery Items: Which of the following are statements of observation?
a. There is a number on this page.
b. A secretary typed this sentence.
c. This page was written after page 4.
d. This statement has seven words in it.

Limits: Correct: All statements that describe something visible to the observer.
Incorrect: All statements describing something not directly visible to the observer, but readily inferred from visible objects or events.

Entry Skills: To identify descriptions of what one sees." (30)

If you have specific instructional plans, your number of tryouts is likely to be reduced. By simply reviewing the plans a great many expensive errors can be picked up early.

You can conduct constructive evaluation on units or segments of a total project. In a short time, by studying instructional plans, characteristics that pertain to the construction of more than one unit can be discovered; you might, for example, find that certain kinds of action sequences attract viewers in each of a series of films. Thus, evaluation results on one unit may have implications for many units. Later, because of the detailed plans and unit commonalities you may be able to pinpoint that these factors of instruction contribute to success or failure.

In your instructional plans you might specify the medium to be used -- text or television, for example. There are advantages to this. By knowing the medium, you can explore its limits. You can build tests to check those limits more precisely and given enough resources early, you can compare media for their ability to achieve a set of objectives.

CASE

Stating Instructional Plans

Health Show researchers created a notebook including instructional plans for producers and writers. To produce the health show, specifications of methods and material had to be made for writers, as a great deal of the show's content was likely to be complex and unfamiliar to writers and producers. Researchers made an outline for the health show's producers and writers' notebook based on what was learned from interviews, task force meetings, and library research. In addition, they included valid principles of behavior change and influence strategies to help writers and producers to achieve the greatest effect. The outline below is presented just as it was when suggested by researchers. Researchers were supposed to fill in the blanks for each topic area.

DRAFT

OUTLINE OF POSSIBLE CATEGORIES FOR THE HEALTH SHOW PRODUCERS' NOTEBOOK (31)

I. Overall Goals

  A. Information Goal

    1.

    .

    . Specific informational details

    .

    x

  B. Attitudinal Coals

    1.

    .

    . Specific informational details

    .

    x

  C. Action Goals

    1.

    .

    . Specific informational details

    .

    x

II. Thematic Corollaries

  A. Major Themes

    1. Theme A (e.g., you can do something about it yourself)

    2. Theme B (e.g., you have a right to good health)

3. Theme C (e.g., the positive side of good health)

.

.

.

x

B. Topic -- or goal-specific themes

1.

.

.

.

III. Target Audience Corollaries

A. Target Audience

1. Demographic characteristics as they relate to the treatment of the goal and ease of implementing the goal.

2. Cultural characteristics " " " " "

3. Attitude characteristics " " " " "

4. Habit characteristics " " " " "

IV. Possible Influence Strategies

A. The information step

facts, statistics, concepts, principles, related to motivation for positive action

B. The motivation step

1. discuss attitudinal and motivational factors

a. as instrumental goals

b. as possible obstacles

C. The action step

discuss or display action strategies, imitative models, possible habits to be developed or modified

D. The relay step

1. deal with how some viewers can affect others (for example, two-step flow and peer-influence approaches)

V. Caveats

A. Myths and Misconceptions

B. What the producer should avoid in treating a particular goal or topic

From the contents of the health show producers' notebook, C.T.W. researchers will be able to ask more precise evaluation questions: "How do certain specific information details, their amount and complexity, affect a viewer's ability to remember the information? How do people react to a specific form of an action step (IV B)? Do people follow the request to relay the information to another (IV D)? What effects would it have on a person's behavior if he were shown a model of someone relaying information?"

Precise description of instructional plans will help you to ask more exacting questions. Your specification of content, combined with your instructional plans, will help you to ask questions about the precise theoretical links between your course and your goals. This is an early check on the consistency and probable success of all instructional components -- goals, content, tests, and instruction.

* * * * * *

Do you have sufficient resources?

After you have decided that constructive evaluation is the appropriate process for your project, you must determine if you are ready to begin planning the evaluation. First, as described in the last section, you take an inventory of the ideas which describe your course; then you account for your resources.

Without the necessary resources, the attempt at constructive evaluation would be self-defeating and frustrating. Resources include available time, money, staff, tryout groups, facilities, equipment, test systems, early drafts of materials, and locations in which to test the groups. (32) A relatively large percentage of each resource allocated to a project is necessary for constructive evaluation and proper revision.

If you have enough time to collect and use data, you have enough time to begin to plan a constructive evaluation.

You will need time to answer constructive evaluation questions. In programmed instruction, for example, approximately thirty-six hours of development, some of which is evaluation time, are necessary to produce one instructional hour; in computer-assisted instruction the rate can be up to 400 development hours to one instructional hour. (33) In a typical classroom setting it may take a teacher one year to show improvement depending on the problem. (34) Thus, you can imagine what it would take for you to plan, test, and revise a large-scale project. But evaluation time may be shortened when (other factors being equal) you have a set of tests available, you have a sample of people available, (35) and you can turn out a rough unit quickly.

You must find out the rate and amount of time allotted for production of materials, and the amount of time allotted for testing, revising, and testing again. This is necessary to assess your time limitations and the possibility of coordinating your evaluation with your production. The more time you allot for planning, in proportion to the time you allot for production and evaluation, the more likely your system is to succeed. (36)

CASE

Planning Time to Collect and Use Data

The health show staff had a period of eight months to produce and test a pilot. Most of this time was used to plan the pilot videotape. During the last three months, when the pilot was produced, tests and tryout procedures were perfected on other similar films, and a major test of the pilot was planned and conducted. Following the pilot test, producers and researchers spent their time analyzing the data and deciding upon what to revise and what to keep. After the pilot, the producers had a year to produce the first season's shows.

> If you have money allotted for collection of data and revision of instruction, you have enough money to begin planning a constructive evaluation.

It is difficult to compute the exact dollar allotment for constructive evaluation from the budget of the health show project: constructive evaluation is so well integrated into C.T.W.'s operations that the functions, roles, and the costs of production and research overlap. I can say that the usual allotment of money for constructive evaluation is between 5 and 20 percent of the total budget, depending on the size of the budget. Thus, for a $200,000.00 project you might allocate $10,000.00, or 5%, for constructive evaluation. For a $200.00 project you might set aside 10-20%: $20.00 - $40.00. (The budget for the whole second year of production and research for the health show was seven million dollars.)

Be sure that the evaluation budget for your project shows an amount specifically set for constructive evaluation. Many funding agencies ask project directors to devote a relatively large sum of money to constructive evaluation.

You will be able to decide on this budget by considering the sources of expense which may include:

a. Developing new tests.
b. Instructing the classroom teachers who are to cooperate by trying out the method or product.
c. Buying and maintaining equipment.
d. Teaching your staff the skills of measurement and observation used in tryouts. (37)
e. Paying experts to review materials. (38)
f. Preparing materials for use in tryouts.
g. Recruiting manpower (39) with reserve workers. Most manpower costs accumulate during planning, designing, and creation of rough drafts, with the heaviest costs appearing in the planning and draft phases.
h. Diverting learner time. (40) In industry, time spent away from work for training purposes costs money; in schools subjects used in a tryout are sometimes paid.
i. Purchasing or renting of a test environment.
j. Expending miscellaneous funds. It may be difficult to pinpoint some costs because the same money may be put to multiple uses. (41)

The project must be funded well enough to finance the making of a number of extensive and sometimes expensive revisions. (42) For expensive instructional methods, revisions resulting from constructive evaluation can double the cost of the development phase of an untested instructional system. Thus, a good rule of thumb is to get twice the funding that you think you need for one draft, and, if you can, include a clause in the proposal that will enable you to get more money for evaluation if necessary.

You will arrive at the amount of your evaluation budget by the importance you attribute to the desired results and the degree of experimentation in the project. The more important the results, the more likely it will be that you will have large amounts of funds allotted for evaluation. The more advanced and original the project, the more uncertain you will be in your predictions of the results, and the greater the amount of evaluation money you will need. The more uncertain the outcome, the less confidence you should place in the cost estimate. (43)

You should ask and answer cost-effectiveness questions early in the process; that is, estimate the likely results and decide what you are willing to spend to get them. You may begin to compute cost effectiveness by determining the costs of alternative methods to reach goals.

But cost effectiveness need not compare two courses. You can compare two aspects of the worth of a single program: 1) you can estimate the usefulness of the instruction (amount of learning, time saved, number of students served), including the estimated reliability of achieving its results and 2) you can compute the costs necessary for producing the method or product. (44) When analyzing only one system, you may find two types of acceptable results: 1) you may find that a program costs less than anticipated and its effectiveness is greater than expected or 2) you may find that costs are less than you thought and effects are about the same as predicted. (45)

A course can be useful in many ways. The cost of a program which accounts for a few moments of students' time can be pro-rated in terms of thousands of potential students making use of the successful system. For example, when the cost of making 130 shows per season for "Sesame Street" is divided by its nine million viewers, the cost is about one-third of a penny per child per program! (46) You can also include as useful results such by-products (47) as sophistication of users and producers, operating rules, design principles, further knowledge of the subject matter, population sophistication, and incidental outcomes.

In any case, money is essential for a successful constructive evaluation of a large-scale project. If you do not have enough money, you may be giving up a certain quantity and a certain quality of student learning. Be sure you have enough.

> If you have a well-trained staff, you have a valuable resource to help you begin planning your constructive evaluation.

You can consider your staff as both resources and limitations: their skills and training are resources; their past habits of instructional development are likely to be limitations. Therefore, you must select qualified staff members or you must train them.

CASE

Selecting and Training Staff

In the early stages of the health show, the staff included an executive producer, an assistant production director, a writer, a technical expert, a research director, and several researchers and secretaries. Each staff member had considerable experience in his own field. For example, the technical expert was a Ph.D. in biophysics, the executive director had produced many shows, the research director had worked in other C.T.W. research projects, and one staff researcher was an expert at survey research. But the producers were not well enough versed in the uses of constructive evaluation in the production of instructional television. They had to be taught how to use the data gathered.

> If you can fit your constructive evaluation into institutional constraints, you can begin to conduct your evaluation easily.

You should be aware of your institution's operating procedures so that you can easily fit your ideas of evaluation into the process. You must learn how the institution operates, and you should learn the purpose and function of the institution, its facilities, equipment and staff, and you should identify those people inside the institution likely to affect your program. You must also know to whom in the institution any information should go and those in the institution who make the decisions.

To be able to work best within the operating procedures of an institution, you must gain the administration's support. You should be sure you will be free from administrative pressure that will bias your results, change standards, or reduce the quality of the system. You should also be free from staff pressure to gloss over first mistakes: production staff may be afraid of looking bad; the constructive evaluation will reveal their faults.

You must plan the initial application of a constructive evaluation in an institution so that the evaluation system will survive. (48) To survive beyond its first use, the process must be an essential part of the institution's total instructional development system. A good survival plan for constructive evaluation on a major project should include a permanent budget, permanent staff, and permanent space. You can increase the probability of the continued use of evaluation by providing information useful to the institution. To aid the survival of the constructive evaluation process, its procedures, and measures should be useful in more than one instructional lesson or project. Finally, to be most effective, any change in the constructive evaluation procedures should be purposeful and based on results, not institutional politics or lack of use.

CASE

Working Within Institutional Limits

Health show producers and researchers studied the development of other C.T.W. projects: "Sesame Street" and "The Electric Company." They used the workshop's approach to deriving goals and objects by the use of task force meetings. They identified people in other sections of the Workshop who could affect their work, like those people in the Community Relations Department who could help get people for task forces, who could describe their target population, and who could arrange for groups to test. They determined which individuals in the administration should hear about important decisions as they were made. But according to institutional policy, producers were given control over the major decisions regarding their show.

Constructive evaluation staff members were hired and were settled in their office spaces. As is usual in C.T.W., constructive evaluation was to continue well beyond pilot testing.

Several groups and institutions which would affect the production and use of a health show were considered by C.T.W. staff. The resources of the Public Broadcasting Service for example, was taken into account:

> "...a weekly series of 26 hour-long programs to be broadcast, first, in the early evening over the 240 stations of the Public Broadcasting Service. We expect that most stations would repeat each week, and at a different hour, to reach the widest possible audience. We also expect that the programs would be broadcast during high school class time, so teachers could make it part of their health education courses. In addition, we believe that stations would occasionally broadcast local "follow-up" programs like those that helped give (other successful shows like) VD Blues its remarkable impact." (49)

If you know about your potential students, and take that knowledge into account, you will be able to complete your evaluation.

Information about potential students -- those people most likely to need or want to learn from the method or product being developed -- should be used during the early stages of planning. You will know from students' accessibility and location how likely you are to be able to complete your evaluation successfully. For example, if for an audience assessment certain students chosen for their age and socio-economic status showed, via pre-test, little of the skill or knowledge to be taught in the program, but revealed a positive attitude toward the subject, you would have an ideal group of subjects: untrained and cooperative. (50)

To collect audience information, you can gather data from existing sources if it is available. With permission, you can even enter into real environments and homes to collect evidence of preferences and opinions by interview and observation. (51) Audience prerequisites can be obtained by seeking expert advice, checking existing data, or exploring through observation. In brief, you can get student information in many ways:

1. Written records
2. Interviews
3. Questionnaires
4. Data from the real environment.

You may look for anything which you believe may influence their ability to learn from the instructional program: ability to read, attitude toward subject, age, and present habits. (52) (53) (54) (55)

With knowledge of potential students, you can construct pre-tests and final tests for a unit. For example, in some early research done for an instructional program to teach Spanish-speaking and English-speaking children the vocabulary of reading instruction, student knowledge was carefully assessed. (56) Key concepts were chosen after testing children and analyzing the instructions used to teach reading in teachers' manuals. Students showed they did not know 17 words often used in teachers' statements; for example:

> "Direct the children to look at the picture at the top of the page..."
>
> "Then direct them to find the picture at the bottom..."
>
> "Mark the word in each box that is different..."
>
> "Color the pictures that are alike..."

Instructional developers chose from the 17 words to create the content of the unit, and they constructed the pre-test and the final test of the unit to verify mastery of these concepts.

CASE

Choosing an Audience

The health show staff decided upon the audience for the show. The family was to be the target; minority group families and poor families, urban and rural, were to be given special consideration. From interviews with experts, the staff found that the American family needed to learn much about health knowledge and practice, and that the American family members had to be motivated and convinced if they were to take an interest in protecting their health.

Through the many personal and business contacts of Vivian Riley of Community Relations at C.T.W., and C.T.W. affiliations with community centers around the nation, many people from the target audience could be found. They were accessible, but were they controllable? They could be found anywhere in the country, but to reach them, procedures to test the program had to be devised for home viewing or community center viewing.

The health needs of the people in the community were carefully studied to decide upon the best audience. The health show producers and evaluators tentatively decided that:

> "...although the health series will be aimed at all adults and teenagers, we are convinced that its

> primary target audience should be mothers and young parents, and that the needs of the poor and lower socio-economic groups should be of special concern. Dr. Shervert Frazier, Chief of Psychiatry at McLean Hospital, Belmont, Mass., told us: 'Mothers are wholly responsible for family health, in crisis as well as routine matters.'"
>
> "Basically however, we conceive of the series as a family service to be watched by the entire family unit." (57)

If evidence of the effectiveness of your instructional method has been collected by others, you can save considerable time and money in your constructive evaluation.

You can save a great deal of time and energy if your library research, and other pre-production studies, reveal information which has a bearing on the effectiveness of the instructional methods or materials you are to use. You might find enough evidence to curtail the extent of your constructive evaluation.

CASE

Searching for Existing Evidence

The health show staff studied whatever data was available on hundreds of health commercials, films, and television programs. They learned about the successes of television programs such as "V.D. Blues" and the relative failures of some anti-smoking commercials. They saved a large amount of time and money by not having to rediscover what was known.

If you can find an existing test for determining the effectiveness of your system, you will save the time and money required for creating one.

You can be most efficient if, at the beginning of the process of constructive evaluation, you find as many tests as you can that answer your questions and assess the ability of your students to perform according to the goals you wrote. Once you have collected some tests, you should check them against the criteria given in the chapter in this book on tests used for constructive evaluation.

CASE

Perfecting an Existing Test

Dr. Keith Mielke, a professor on leave from the University of Indiana, working for the health show, found some existing test procedures which fit some of the questions asked. The health show staff wanted to know if the show would be appealing, so Dr. Mielke found an existing test procedure suitable for determining appeal:

> "A variation of the apparatus known as the Stanton-Lazarsfeld Program Analyzer is being employed for moment-by-moment measures of appeal."
>
> "Each subject has two push-button switches, a red one for the left hand (for registering 'dislike'), a green one for the right hand (for registering 'like'). Without undue disruption of the ongoing program, a group of (10) persons can 'vote' repeatedly simply by pushing one button or the other. The votes are recorded on a moving paper scroll in the event recorder, allowing after-the-fact matching of votes with precise program content. For the type of programming tested so far and anticipated in the future, a voting interval of once per minute (50 seconds off, 10 seconds on) seems to be working well. Respondents are cued that they are to vote by a red light on top of the TV receiver that flashes on and remains on for the ten-second duration of the voting period. When the 'voting light' first comes on, a soft tone sounds also. The voting intervals are controlled by automatic timers that can be set for various intervals. A manual override feature allows the researcher to call for a vote at any time in the program whether or not this falls in the ten-second regular voting interval. The cumulative tally of votes for 'like' and 'dislike' are plotted across a time line for easy interpretation." (58)

If you try your test procedures on existing materials while your own materials are being produced, your constructive evaluation will be more likely to start on schedule and proceed easily.

You need not wait to produce materials to get some feedback on your teaching ideas. While waiting to finish a prototype, you can be testing existing materials which are similar to the instructional method you have in mind.

CASE

Trying Tests on Existing Materials

The health show staff reviewed and evaluated existing television programs and films. Before the pilot materials (sample reels) were ready, a few existing programs, which were recently done and well produced, were selected for more intensive study, to begin to answer evaluation questions. The programs had to be appropriate for the bulk of the target audience, and had to be strongly related to the topics or techniques to be used in the sample reel.

Two of the existing films selected were "I am Joe's Heart" and "V.D. Blues." The reasons for selection follow:

> "One program selected for intensive testing was the Reader's Digest film entitled I am Joe's Heart. Although the production was highly stylized and somewhat wordy, its treatment of the subject of heart attacks, after the opening five minutes, was not lecturing in tone, and it seemed to have potential for interesting at least a narrow spectrum of the audience in the subject." (59)
>
> "Another program selected for intensive testing was a PBS Special of the Week entitled VD Blues. In many respects, this program came the closest among the existing materials to approximating some structural elements of the anticipated health show: it was entertaining; it addressed a significant health issue; it incorporated a variety of production formats within a type of magazine show, albeit a single issue magazine show; and it was distributed through the Public Broadcasting System." (60)

* * * * * *

Are You Committed to Revision?

Even if you locate the resources to test and revise, your whole process of constructive evaluation may be a waste of time, energy, and money if you make no commitment to revise. To be committed, you and your staff must see the project as important: you must view student learning as the desired result.

Any behavior which shows that you and your staff are willing to give time, energy, or money for collecting and using information to improve your program is evidence of commitment. If you ask questions, allocate staff time, spend time in meetings to consider tests of the project, ask for meetings to help coordinate testing with production schedules, then you have shown evidence of commitment.

Your staff members' early behaviors may be promising, but the real test of commitment is their consideration of the data collected. To be sure of your staff members' reactions, you can ask them to turn out a short prototype and test it immediately. This will provide you with an opportunity to see if they will consider data when revising.

CASE

Demonstrating Commitment

The health show producer's commitment to use constructive evaluation for revision could be seen by his allotment of time and money, his questions, and his overt statements, which referred to revisions and changes of the pilot to be made. In addition to the allotment of staff, funds, space, and equipment, the production staff seemed to be committed to using the test information gathered on the pilot to improve the instruction as it developed. They stated:

> "For the first time, such a project will be based on the continuing guidance of expert advisors."

> "All entertainment elements in the sample reel (the pilot video tape) are subject to change if investigation should find them inadequately diverting or insufficient as an 'influence strategy'." (61)

* * * * * *

Summary

To make the decision to go ahead with constructive evaluation and to form evaluation questions, you check your resources in terms of time, money, staff, students, tests, procedures, and materials, and your specific ideas for instruction, considering goals, content, theory, and instructional plans. The more specific your ideas the easier it will be for you to formulate your evaluation questions. The earlier you allocate your resources and express your ideas, the easier the constructive evaluation will be, and the greater the rewards from the information collected.

CASE

Beginning the Process

When they had enough ideas to form evaluation questions, and once they were reassured by the commitment of resources and by their own willingness to use the information to improve the show, health show staff members decided to forge ahead with the full process of constructive evaluation.

* * * * * *

Determining the Need for Constructive Evaluation, in Brief

Deciding: Should I use constructive evaluation?

- Do I need information that will guide my work while creating this project to...

  ...resolve doubts
  ...make improvements
  ...contribute to instructional theory
  ...save money

Beginning: Should planning begin?

- Do I have specific ideas about the instruction? Do I have...

  ...goals
  ...specific course content
  ...theoretical principles
  ...instructional plan

- Do I have sufficient resources? Do I have...

  ...time
  ...money
  ...well-trained staff
  ...information about potential students
  ...information from research literature
  ...existing tests
  ...existing materials

- Do I have commitment to revision?

CHAPTER IV

Writing the Recipe: Forming Evaluation Questions

When making decisions in constructive evaluation, you will return again to your evaluation questions, just as a chef returns to a recipe while cooking. Your questions are the focal point for the rest of your constructive evaluation; you collect information to answer the questions you form, and you revise based on the information you collect. Therefore, when forming your questions, you should be comprehensive. Include all those ingredients which influence your course results: your instructional method, your audience, and your instructional setting.

CASE

Asking an Evaluation Question

> The Children's Television Workshop created "The Electric Company," a television show produced to teach reading to third and fourth grade children who had not learned to read well. The main constructive evaluation question asked by the producer was: "Will we be able to improve our program as it is being developed so that it teaches slow-reading, urban and rural children to be able to read by use of an entertaining half hour, magazine-format, television show to be broadcast into homes and schools?" The method, audience, and instructional setting were all included in the question. (1)

> Constructive evaluation questions may include instructional method, audience, setting, and results.

Your questions may have many facets. You may wish, for example, to inquire about effectiveness, efficiency, and acceptability. Or you may be concerned about certain methodological features of your course or you may be interested in audience characteristics and their effects on results. Perhaps you want to be sure that course procedures go as planned: that events take place as described, that the audience is like the one requested, and that the setting is according to specifications. And you will probably want to know why results come out the way they do; you may wish to seek clues from the interaction among method, audience, and setting.

CASE

Expanding the Question

> When the producers and researchers of "The Electric Company" thought about the problems of creating a show that would truly help children learn to read, they formed many questions:
>
> > Will children learn to read? Will slow readers catch up? Will illiterates learn to read at all? Will they learn the sight words? Will they learn to apply the phonics rules? Will the amount learned from the show (and associated methods and materials) be worth the cost? Will the slow readers pay attention to some aspects of the magazine format and not others? Will the show maintain an individual child's attention when he watches it with a classroom full of other children who react to the show? Will a one-half hour show presented five days a week be sufficient to aid children in reaching the goals? Do their eyes scan the words? Do they comprehend the humor? How much can a child learn about reading in one-half hour? Do they comprehend what happens during the show? (2)

> The questions you ask will be determined by their importance, their consistency, and their financial feasibility.

You will probably generate more questions than you could ever answer. You will have to select the most important and the most feasible ones.

Choose questions which can be answered by an evaluation which you can support with your available time, money, staff, tests, and audience, and select those questions which include important results which you believe may not be easily achieved by your program. Select questions which may reveal program faults which you could afford to revise, and those questions which are internally consistent: the method mentioned in your question is likely to lead to the results you desire, given your audience and setting.

CASE

Selecting Questions to Answer

> Not all of the questions asked by producers of "The Electric Company" could be answered, at least they could not all be answered during the first couple of years of production. The producers did not have the time, money or staff to answer them all. So they picked those questions they felt were internally consistent; those in which the method seemed to lead to the results with their audience. From these questions they picked those which mentioned results which they felt could not be accomplished easily: there were some strong doubts. Of those questions, they picked the ones which they could afford to find out about. The questions they decided to explore were:
>
> Will slow readers pay attention to some aspects of the show and not others?
> How much can a child learn about reading in one-half hour?
> Do children comprehend what happens during the show? (3)

* * * * *

Summary

Evaluation questions provide the basic recipe for the rest of constructive evaluation. You define each part of the question and then choose a way of representing each ingredient for a test of your project. Your methods are represented by drafts of materials, your audience is represented by a sample of people, your setting is represented by the place where the tryout is conducted; your results are gauged by tests. It is in this way that evaluation questions provide you with the elements for the rest of your constructive evaluation.

* * * * *

Forming Evaluation Questions, in Brief

Write questions including...

- ...instructional method.
- ...audience.
- ...setting.
- ...results.

Choose questions which are...

- ...important.
- ...consistent.
- ...feasible.

CHAPTER V

The Bridge: Defining Elements in an Evaluation Question

Consider this evaluation question related to the television production called "The Electric Company": "Will slow readers learn to read from an entertaining, magazine-format, television show broadcast to homes and schools?" Can you plan a test of the program "The Electric Company" based on the mere mention of each instructional element in the evaluation question? The answer should be "No"! The elements included in this question are too vague to use for choosing a test, a sample of the audience, a test site, and a prototype representing the method.

Why should you choose and define each element? Because you need a bridge between an evaluation question and a test of a project: the more precisely you define an element, the easier it will be for you to make the transition from a question to a test, a prototype, a test site, and a sample audience.

> If your goals are specified, you can choose or create tests and estimate the effects of the program in achieving those goals.

You simply explain what you mean by the goals mentioned. For example, student attention, a desired result, may be defined as the orientation of a person's face and eyes toward a book, teacher, or T.V. screen, or the ability to repeat a statement immediately after it is made, or the movement of one's eyes across a screen. These definitions describe the result "attention" in such a way as to suggest what to look for and how to look for it.

If you describe your audience's characteristics, you will be able to choose a sample audience and possibly account for the effects of the audience on the program.

An audience for a math teachers' training course may be defined: "Any teacher responsible for teaching math, who knows basic math principles, but who has not had previous training in use of cuisennaire rods, and who is willing to try different techniques of teaching math in his classroom." The audience, "math teachers", is defined in such a way as to enable an evaluator to accurately select a group of people for a test of the training course.

If you describe the features of a particular instructional setting, you will find it easier to choose the place for a tryout and infer what effect the environment may have on the program's results.

The instructional setting for an in-service course on management, to be used at an automobile plant, may be defined: "Any room which provides adequate, controlled lighting and temperature, comfortable, movable seating at desks for twenty students facing a screen, a carousel slide projector, and a tape player suitable for group listening." The setting, "the room," is defined in such a way as to direct an evaluator to select a location with the appropriate characteristics.

> If you specify the characteristics of your instructional methods and materials, you can choose an instructional unit for use in a tryout.

A method of teaching language skills may be defined: "A module consists of a) a programmed teacher's manual which instructs a teacher how to use linguistic principles and reinforcement techniques; b) a series of booklets and tapes, three per unit, including objectives, linguistic exercises, and cases to solve, practice tests, and answers, and c) a kit to make language learning materials for individual student use or for class demonstrations which will carry out ideas in the teacher's manual." The method and materials, "language modules," are defined in such a way as to help an evaluator choose a representative instructional unit for a test of the method.

> Define elements in your question by observing your intended audience interacting with an early draft or the completed version of a similar instructional method.

CASE

Defining a Question Using a Similar Product

A one-hour television show, "V.D. Blues," was broadcast on the public broadcasting channel in most states for a week or two during 1973. The show used a magazine format and included songs, sketches, film of live sequences, and straight information presentation. It was produced for the purpose of informing adults and young people about the causes, signs, symptoms, and effects of venereal disease. Two of its main objectives were to inform people where to get a test for V.D. and to get the test. In New York, a telethon followed the show. If an audience member cared to find out where he could be examined, he could call in. He could ask anything about the topic he wished.

The health show personnel at the Children's Television Workshop were interested in observing audience reaction to the show because the program was generally like the show they planned. By viewing a method similar to their own, and observing the effects, many questions were suggested and many elements began to be defined. They found variables of interest merely by watching the telethon. The health show staff agreed that phone calls and clinic visits following the program would be an excellent expression of the desired result.

The C.T.W. health show producers also conducted an informal survey of staff members and others who watched the show and asked them what they liked and what they remembered. In this way they began to ask the right questions to define desirable show characteristics: Was a sketch taking place in the uterus in poor taste? What made it seem that way? Did they remember the words to songs? What made certain songs memorable? They became interested in what made the show believable and what did the opposite. Were the actors playing the nurse and doctor who told about V.D. symptoms believable? What made them believable? They were interested in defining the features of the show that seemed to set a serious yet entertaining mood. For example, Dick Cavett, the host of the show set the tone of the show with a few phrases: "Don't give a dose to the one you love most," and "V.D., the gift that keeps giving."

Create specific but flexible definitions.

Your definitions should be specific enough to suggest tests, materials, settings, and sample audiences, but they should also be general enough to give you some room to maneuver.

CASE

Defining a Variable Specifically and Leaving Room for Change

Milton Chen, a student researcher from Harvard, defined one of the aspects of the process of learning as different forms of overt verbal response. His definition was both specific enough to define categories for an observational measure, and general enough for him to be able to make changes.

Chen was looking for a set of behaviors that would indicate that a student was on his way toward learning to read from "The Electric Company." Tentatively, he defined this element by six categories of overt verbal response which would lend themselves to measurement by observation. (Children who don't respond aloud are learning too; Chen was just looking at one overt observable form of behavior which might indicate progress.) (1)

CATEGORIES OF VERBAL BEHAVIOR
EXHIBITED DURING VIEWING OF "THE ELECTRIC COMPANY"

Instructionally-Relevant Verbalization

1. Reading of print on the screen: The child reads or attempts to read print appearing on the screen, regardless of the timing of the voice-over (an unseen narrator).
2. Spoken Anticipation of Print to Appear on Screen: The child pronounces the word in anticipation of its appearance on the screen.
3. Instruction-Related Verbalization of Print: The child comments about print appearing on screen, but does not proceed to pronounce it (e.g., "that word begins with a g," or "That word has an 'oo' sound.").
4. Story-Related Verbalization of Non-Printed Speech: The child verbalizes about plot, characters, setting, attractiveness of bit; or, he imitates the speech of characters.
5. Oral Participation in Songs: The child sings along with all or portions of a song.

Irrelevant Verbalization

6. Other-Than-Program-Related Verbalization: The child verbalizes in a manner unrelated to the instructional message of the program, i.e., comments directed to the program, i.e., comments directed toward other viewers and unrelated to "The Electric Company."

Chen, observing children watching the show and using the categories he defined, found that...

> "The behaviors described in Category 2, "Spoken Anticipation of Print to Appear on Screen," were found to occur chiefly during "Monoliths" [a monolith appears as it did in the film "2001" and shatters to reveal a word] and "UCLA Band" [a marching band forms a word]. Category 3 did not occur with any frequency, nor was it related to any particular bit. Also, Category 4 was reworded to account for the dominant reaction of trying to predict "what happens next," or demonstrating that one already knows "what is going to happen next." The rest of the categories appear to be fairly appropriate descriptions of their respective classes of behavior.
>
> Category 1, "Reading of Print on the Screen," is certainly the most frequently occurring and educationally significant behavior of those encountered in this study. It is also the category which received the greatest amount of attention in judging a bit for verbal response. Observations indicated that a significant amount of vocalizing printed words could not actually be termed "reading;" much of it was repetition or mimicking of the voice-over (we know this because six- and seven-year olds who could not read were reciting many words on "The Electric Company"). To filter out reading partially from mere imitation of voice-over, Category 1 was split into two subcategories: Recitation of Words in Print (Before Voice-over) and (After Voice-over.)

Then he changed the definitions of the categories: "Reading print on the screen has been expanded to recitation before and after voice-over, and oral participation in songs was eliminated."

CATEGORIES OF VERBAL BEHAVIOR OCCURRING DURING VIEWING OF "THE ELECTRIC COMPANY"

Instructionally Relevant Verbalization	1.	Recitation of Print Before Voice-Over: Viewer pronounces or attempts to pronounce words or letters appearing in print on the screen before voice-over pronunciation of the word.
	2.	Recitation of Print After Voice-Over: Viewer chimes in with or repeats words after voice-over pronunciation of words or letters in print.

	3. Verbal Anticipation of Print About to Appear on Screen: Viewer pronounces word in anticipation of its appearance on screen.
	4. Instruction-Related Verbalization About Print, Exclusive of Attempted Reading: Viewer comments about print on screen, but does not attempt to pronounce it (e.g., "That word begins with a g," or, "That word has an 'oo' sound.")
	5. Story-Related Verbalization of Non-Printed Speech: Viewer comments on plot, characters, setting, or attractiveness of bit; anticipates subsequent events; or repeats the speech (not appearing in print) of characters.
Irrelevant Verbalization	1. Other-Than-Program-Related Verbalization: Viewer comments on concerns unrelated to "The Electric Company," e.g., discussion of friends, other activities.

Thus, Chen was able to further define his observational tool because his original definition of learning during a show was broad enough to leave room for change.

To be comprehensive, define methods, audience, and instructional setting as well as results.

Usually most effort is given to defining results. But by the time goals are defined, methods should be fully defined also. To arrive at a definition, you might ask a producer why each feature of the method is necessary to achieve the instructional goals. He might reply, for example, that he includes in his method positive models for students to imitate and numerous examples to make the program appealing.

CASE

Defining an Instructional Method

The Lesson Format described below is the definition of the Kindergarten Art Program developed by Southwest Regional Laboratory. It illustrates a defined instructional method.

> "Each of the sixty KAP (kindergarten art program) lessons will have four components: (1) an illustrated story accompanied by 'demonstrators' -- art work especially prepared to illustrate selected art concepts, and reproductions of masterworks and children's art; (2) a description of student tasks related to the instructional outcomes of the lesson; (3) step-by-step procedures for administering the program and (4) evaluation procedures. In addition, the materials needed for each lesson will be specified; and where these materials are not commonly available in kindergarten classrooms, they will be provided by the KAP. Lessons will be designed for a thirty-minute class period.
>
> Most lessons will be introduced with an illustrated story to be read to the children by the teacher. The stories will be built around an art element/art principle concept, e.g., variety of line. The children will have the opportunity to practice identifying instances of the concept first in the story illustration and, later, in examples of the art work. Explicit instructions as to how to present these materials will be given to the teacher. Discussion questions and expected learner responses will be written on the back of each illustration and reproduction.
>
> Teachers will be given a description of the student task or activity for each lesson. This activity will follow the reading of the story, and, like the story, will be directly related to the learning outcomes of that lesson.
>
> Suggestions to the teacher for monitoring the student's progress will also be included. Student responses, to be reinforced or not reinforced, will be described and, if appropriate, illustrated." (2)

You should organize your goals so you can be certain you have defined the full range of possible results.

CASE 1

Defining the Full Range of Results

Bilingual Children's Television was formed to create a television series addressed to Spanish-speaking and English-speaking children in the United States. Their goals included many mental and social abilities. They wanted Hispanic and Anglo children to learn about each other's culture and language, and feel some pride in their own heritage. The B.C.T.V. staff defined many goals. (3) A few follow:

"1. Sensorimotor: Ability to coordinate a part of the body in a movement to produce a desired effect.

2. Labeling: Ability to identify an object or set of objects correctly by name.

3. Patterning: Ability to recognize or identify the properties of an object.

4. Attribution: Ability to recognize or identify the properties of an object.

5. Classification: Ability to group a set of items on the basis of one or more properties.

6. Combining: Ability to create a new whole by uniting two or more discrete and independent elements.

7. Two-Term Relations: Ability to relate two items along one dimension, for purposes of comparison, showing causation and ordering.

8. Multi-Term: Ability to relate more than two items along two or more dimensions concomitantly.

9. Seriation: Ability to order objects in a progressive series according to one dimension so that each object holds its position with respect to both the object that precedes and the object that follows it."

To be certain they would include the full range of possible results in their planning, staff members integrated social

GOAL GRID FROM THE CURRICULUM OF BILINGUAL CHILDREN'S TELEVISION

Cognitive Abilities

		1 SENSORIMOTOR	2 LABELLING	3 PATTERNING	4 ATTRIBUTION	5 CLASSI-FICATION	6 COMBINING	7 2-TERM RELATIONS	8 MULTI-TERM RELATIONS/ SERIATION
CONTENT	Language Development A								
	Reading B								
	Arithmetic C								
	General Concepts D								
	Music E								
	Art F								
	Science-Nature G								
	Social Structure H								

Social Abilities

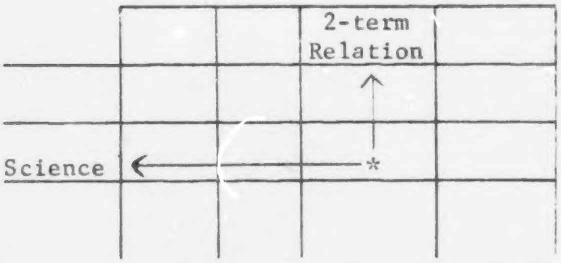
		Appreciating Cultural Styles				
		Within Groups		Between Groups		
		Hispano 1	Non-Hispano 2	Hispano 3	Commonalities 4	Anglo 5
CONTENT	Verbal Communication A					
	Roles B					
	Customs C					
	Diet D					
	Learning Styles E					
	Activities F					
	Environment G					

and mental goals in a grid. (see Chart K) The B.C.T.V. researchers' manual stated:

> "The matrix serves as an analytic framework to insure that the educational goals are constantly considered, and also to facilitate their complete and comprehensive implementation."
>
> "This model insures a maximum of educational richness on the show, as well as preciseness in formative and summative evaluation of show segments."

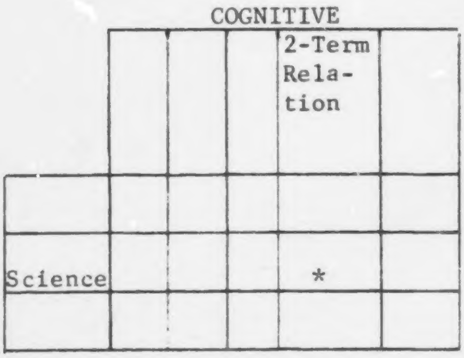
From the cross sections of the grid many more specific results could be defined. The use of the grid is well illustrated in the B.C.T.V. curriculum manual. For example, the researchers stated in the manual that...

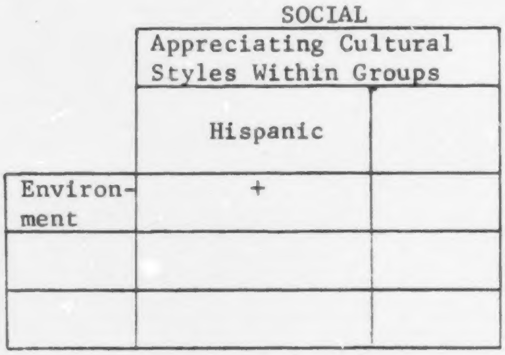
> "a unit teaching the concept that things grow and die, (that is, they change in some way as time goes by,) is represented in a cell thus:


2-term Relation
Science
*

(*) TASK: To recognize that a seed becomes a tree that bears fruit.

> At the top of the column is the reference to the cognitive ability (mental ability) (Two-term relation); at the left is the reference to the content area (Science).
>
> In addition, the above task may also interact with a social unit.


COGNITIVE
2-Term Relation
Science
*


SOCIAL
Appreciating Cultural Styles Within Groups
Hispanic
Environment
+

That is, using the growth of a piño (pine tree native to the Southwest) that bears piñones (pine nuts), as an example to teach the above concepts adds a cultural note appreciating the natural environment of the Chicano population of the Southwest."

CASE 2

Defining the Full Range of Results

To be certain they have defined the full range of possible goals, staff at the Far West Regional Laboratory for Educational Research and Development, used a hierarchical classification scheme to organize goals in a teacher training course. The scheme is shown in the table below:

TABLE: Competency Symbols and Levels (4)

SYMBOL	LEVEL	COMMENT
N.T.	NO TRAINING	
O	ORIENTATION	The task is described or demonstrated; trainee should understand its purpose or function, but cannot perform it.
F	FAMILIARIZATION	Trainee is given practice in performance, but can perform only with close supervision or detailed instruction.
L.P.	LOW PROFICIENCY	Trainee is given repeated practice. He can perform slowly with few gross errors, if given some supervision or adequate job aids.
H.P.	HIGH PROFICIENCY	Trainee can perform efficiently and with no errors. Minimal supervision required.
E	EXPERT	Trainee can teach other people; can invent own solutions. No supervisor required.

When staff members of the F.W.R.L. define a teacher training goal, for example, when they define the goal "to be able to ask questions in class, " they may consider a number of possible results, from orientation, (understanding the purpose of the skill) to expert performance (teaching to others, initiating his own solutions).

CASE 3

Defining the Full Range of Results

The staff of the Southwest Regional Laboratory of Instructional Research and Development defined the full range of possible results for their Kindergarten Art Program by using an art element/art principle grid. (5) Each cell in the grid was a possible result.

S.W.R.L. staff members chose among the possible results by placing X's in the grid according to the appropriateness of the relation between elements and principles.

TABLE - Grid for Identifying Potential Content Areas

ART PRINCIPLES

		balance	dominance	proportion	rhythm	variety
ART ELEMENTS	color		X			X
	line		X		X	X
	texture		X			X
	shape/ form	X	X	X	X	X

Staff members could define a result by combining elements and principles and requiring a child to identify or apply his learning. Here are some of the defined results in the form of behavioral objectives. Consider number 6, in the table below, which is the intersection of the art element "line" and the art principle "variety," and requires a child to apply learning.

TABLE: Unit Objectives for Part of the S.W.R.L. Kindergarten Art Program

"UNIT IV: LINE

1. Given the following types of line, the child will name them: straight, curved, thick, thin, heavy, and light.

2. Given crayons (or paints), the child will draw (paint) the following types of line: straight, curved, thick, thin, heavy, and light.

3. Given a variety of materials for making crayon rubbings, (cotton roping, pipe cleaners, etc.), the child will select and use materials which will vary along the line dimensions of contour, thickness and density so as to produce at least three different types of line on paper.

4. Given construction paper and crayons, the child will tear an abstract shape and give it an identity by drawing in the characteristics of the object suggested by the shape.

5. Given a theme illustrating center of interest, the child will depict the center of interest by using a thick or heavy line.

6. Given a theme illustrating appropriate use of line (e.g., thick telephone poles with thin wires), the child will depict the theme using the appropriate line variations."

> You can be sure you have defined all facets of the elements in your question by analyzing the makeup of each element.

You could analyze elements by creating a sentence which shows the possible variations in each element. The total sentence contains the complete range of variables involved in the definition of an element or a number of elements.

CASE 1

Analyzing a Subject to Define Question Elements

Early in the formation of the Kindergarten Music Program for the Southwest Regional Laboratory, Dick Piper, a project director, formed this sentence relating to his results. (6)

	Musical Elements		Stimulus Mode
A student may be given... (choose one)	...rhythm ...melody ...timbre ...form ...harmony ...expressive elements	in the form of... (choose one)	...singing. ...playing. ...notation. ...verbal. ...body movement.

	Response Mode		Response Type
Students may respond by... (choose one)	...singing ...playing ...notation ...verbal ...body movement	in the form of... (choose one)	...imitation. ...selection-judgment. ...constructive-creative.

He picked different items from each list to form a number of possible results. In this way, Piper could be sure he was addressing those aspects of his evaluation questions in which he was most interested, and that nothing was missing.

CASE 2

Analyzing All Elements in Depth

Audience, method, and setting can be included in the sentence. Lewis Bernstein, a "Sesame Street" researcher, was interested in the relation of "Sesame Street" methods and results. Here is an early draft of one part of a sentence whose purpose was to relate all aspects of the show's evaluation questions. Here you see the parts concerned with one sort of result, attention, and some variables involved in the instruction method. (7)

An Excerpt from an Information Map Which Specifies Variables Related to Attention Results and Educational Methods

...and containing strategies which may serve to catch attention

E_I Catch Attention[*1]

1. Visual channel
  a. Visual mobility (hi-low)
    (1) within frames
    (2) pacing - across frames
  b. Visual density/composition
2. Audio channel
  a. Audio loudness
  b. Verbal mobility
  c. Sound effects
  d. Music
    (1) rhythm
    (2) tempo
    (3) dynamics
3. Cross-Channel Combinations[*2]
  a. Symmetrical presentation (i.e., redundant information)
  b. Complementary presentation (i.e., contrasting information)
4. Repetition
  a. Re-presenting the familiar
    (1) With or without variation

and/or direct attention

E_{II} Direct Attention

1. Visual channel
  a. Action
  b. Incongruity
  c. Salience
    (1) flashing letters
    (2) location of print
    (3) .
    :
  d. Camera techniques
    (1) camera angles, zooms, pans, etc.
    (2) pacing
      (a) changes, variations in scenes, characters
    (3) Special techniques
      (a) e.g., ripple effect
  e. Variation
2. Verbal Channel
  a. Dynamics (loud to soft noise to silence)
  b. Pace (fast to slow)
  c. Incongruity
  d. Salience
    (1) puns
  e. Music
  f. Sound effects
  g. Variation
3. Cross-Channel Combinations
  a. Complementary relationship
4. Explicit invitations to participate

and/or sustain attention

E_{III} Sustain Attention

1. Visual channel
  a. Pace
  b. Density of images
  c. Diversity of
    (1) images
    (2) characters
2. Verbal channel
  a. Rhyme
  b. Accents or dialects
  c. Diversity of voices
3. Cross-Channel Combinations
  a. Complementary presentation
4. Length of segment
5. Story-line
6. Humor
7. Anticipation
8. Explicit invitation to participate

---

*1. The division of attentional strategies into three sub-facets: CATCHING, DIRECTING, & SUSTAINING, is a division used by Gerald Lesser, 1972 (Harvard Educational Review). This division is for the most part arbitrary, and is not meant to exclude the possibility of overlapping orderings, or some other classification.

*2. Lesser, 1972, uses the term "cross-modal reinforcement." This term indicates that reinforcement may not be the only strategy used.

An Excerpt from an Information Map Which Specifies
Variables Related to Attention Results and Educational Methods

F. Educational Strategies

1. Employs direct or indirect method of presentation
   a. Explicit or implicit statement of goal
   b. Explicitly invite the child to participate
   c. Is time (e.g., silence) supplied for reflection?
2. Use repetition
   a. With or without variation
   b. Within the same channel or across modalities
   c. At the end of the segment - e.g., a review
3. Presents goal within a rapid or slow-paced segment
4. Presents goal within a more dense visual image
   a. Amount of clutter in:
      (1) foreground
      (2) background
5. Presents goal within a more dense verbal statement
   a. E.g., number of concepts presented within segment
   b. Number of words presented per minute
   c. Use of verbal puns and rhymes
6. Visual action:
   a. precedes
   b. is simultaneous with
   c. follows

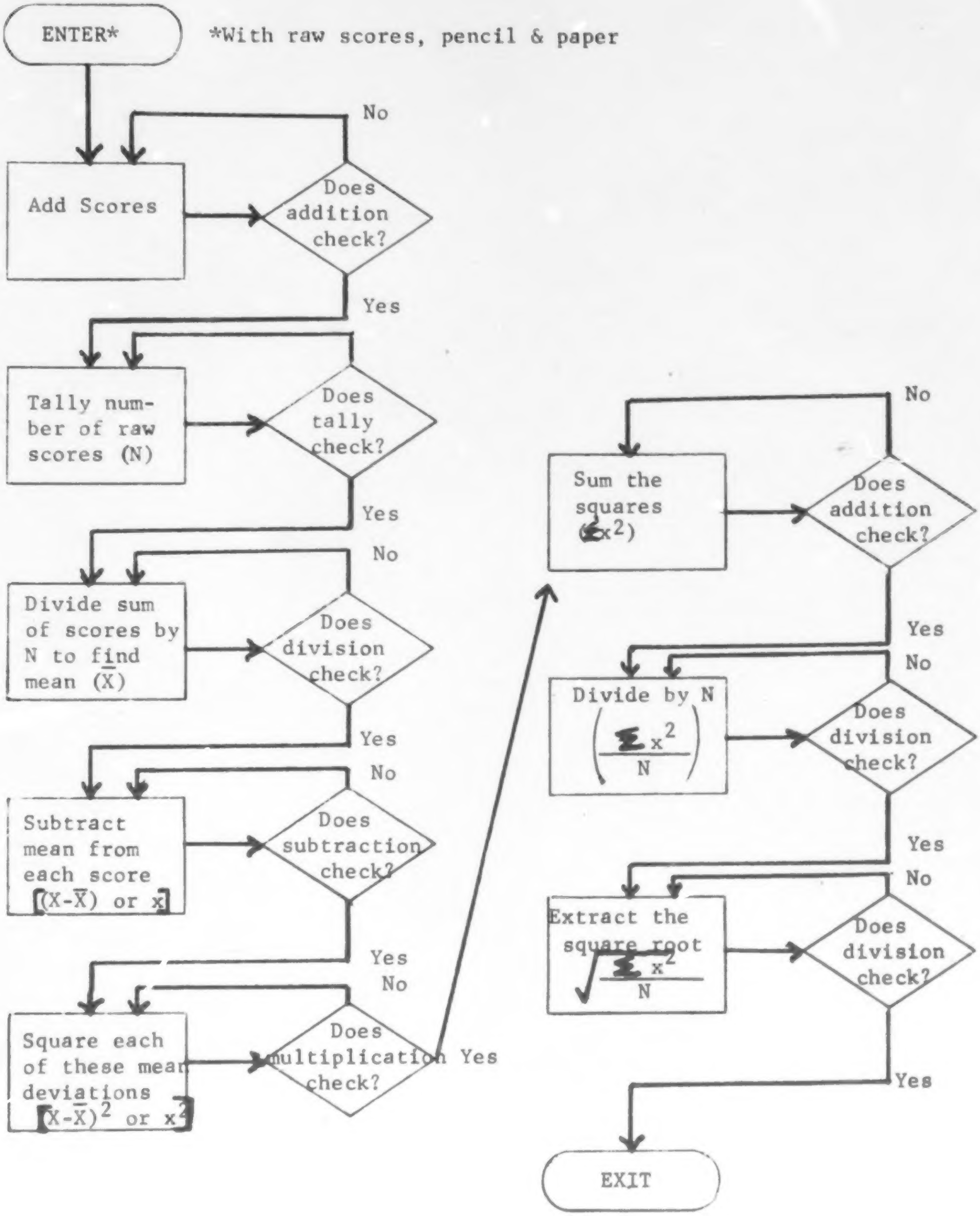
   the appearance of the goal statement
7. Verbal action:
   a. precedes
   b. is simultaneous with
   c. follows

   the statement of the goal.

F. Educational Strategies (Continued)

8. Goal is first stated in verbal/visual channel
9. The goal is made salient by:
   a. format
   b. technique
   c. contrast of alternatives
      (1) in visual channel
      (2) in verbal channel
      (3) in both
   d. through correction
10. a. State a principle and then the example, or
    b. State the example and then the principle
11. The non-educational content relates to the main educational objective (pedagogical integration) -- e.g., the goal is the theme of a storyline.
12. A large-small proportion of time is given to the educational information within the segment.
13. Use the familiar (goal, character, technique) as a bridge to the unfamiliar (Lesser, 1972)
14. Emphasize concrete or abstract representations
15. Present a situation and allude to a conflict:
    a. Show the conflict:
       (1) resolve the conflict
       (2) do not resolve
    b. Do show an alternative to the conflict.

Flow Diagram of the Desired Result - Computing a Standard Deviation


ENTER*
*With raw scores, pencil & paper
Add Scores
Does addition check?
No
Yes
Tally number of raw scores (N)
Does tally check?
No
Yes
Divide sum of scores by N to find mean ($\bar{X}$)
Does division check?
No
Yes
Subtract mean from each score [($X-\bar{X}$) or x]
Does subtraction check?
No
Yes
Square each of these mean deviations [($X-\bar{X})^2$ or x^2]
Does multiplication check?
No
Yes
Sum the squares (Σx^2)
Does addition check?
No
Yes
Divide by N ($\frac{\Sigma x^2}{N}$)
Does division check?
No
Yes
Extract the square root $\sqrt{\frac{\Sigma x^2}{N}}$
Does division check?
No
Yes
EXIT

You can use a diagram representing a task to organize all aspects of an element, and to check for missing portions. In this fashion you can represent your results or you can represent your instructional method. You can study a diagram to be sure that a method has all the necessary content, to see that all the desired results are listed, and all audience prerequisite abilities are considered.

CASE 1

Using a Flow Diagram to Specify Results

If a project director wants to define elements of an evaluation question dealing with a workbook that will teach college sophomores to compute a standard deviation, he might study the flow diagram of the task below:

From the steps and decisions included, he could discover what concepts and skills are needed by the audience, either as prerequisites or as content to be learned in the program (.e.g, raw score, how to check division), and precisely what subskills to look for as resultant learning.

CASE 2

Using a Flow Diagram to Specify Method

Shlomo Waks, an educational technologist at the Technion in Israel, described by flow diagram an instructional method which he used in his doctoral research. (8) By defining his method in this fashion he was able to be sure he had planned all parts of the method for a test of his instructional scheme.

Insert E

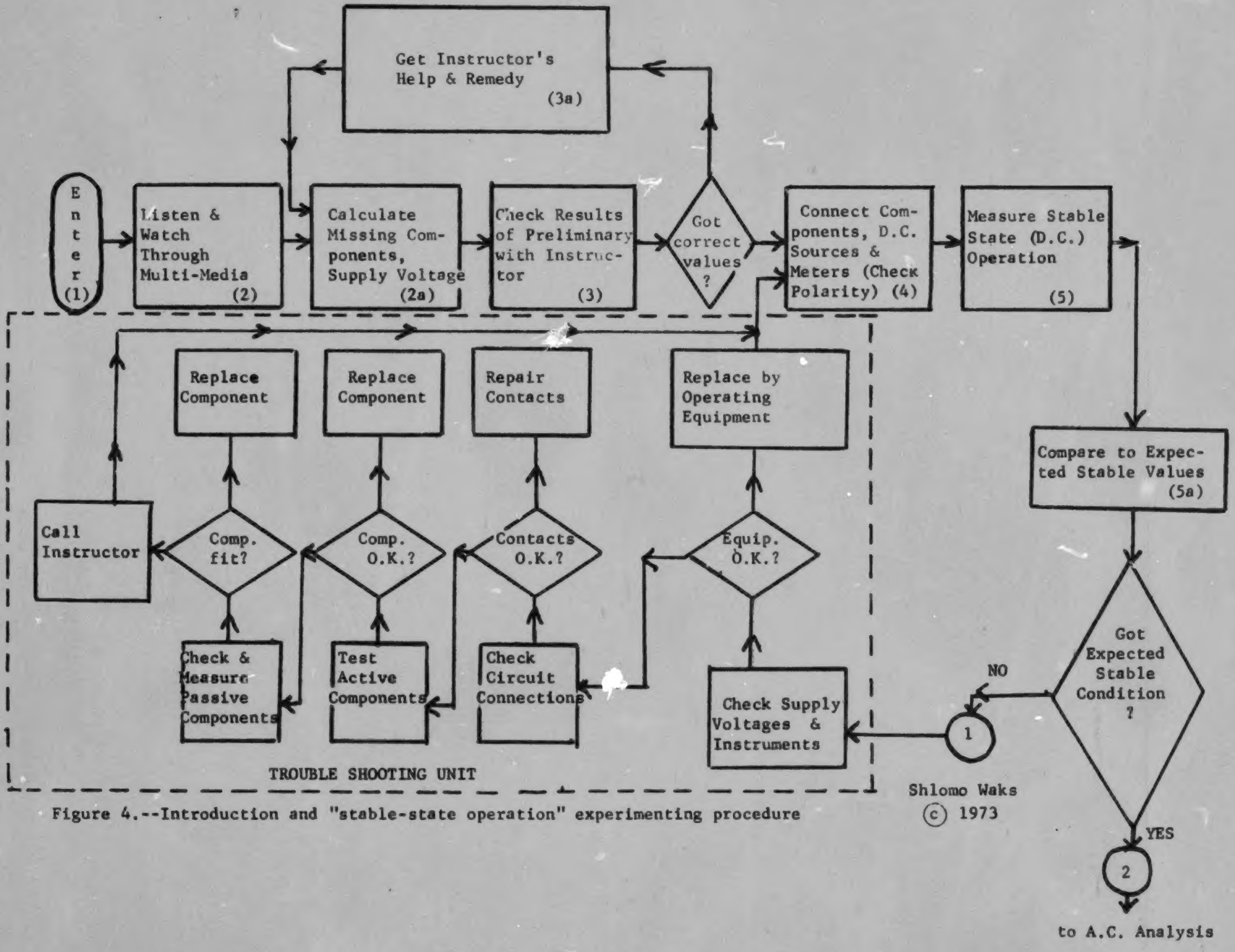

Enter (1)
Listen & Watch Through Multi-Media (2)
Calculate Missing Components, Supply Voltage (2a)
Check Results of Preliminary with Instructor (3)
Got correct values ?
Get Instructor's Help & Remedy (3a)
Connect Components, D.C. Sources & Meters (Check Polarity) (4)
Measure Stable State (D.C.) Operation (5)
Compare to Expected Stable Values (5a)
Got Expected Stable Condition ?
NO
YES
1
2
to A.C. Analysis
Check Supply Voltages & Instruments
Equip. O.K.?
Replace by Operating Equipment
Check Circuit Connections
Contacts O.K.?
Repair Contacts
Test Active Components
Comp. O.K.?
Replace Component
Check & Measure Passive Components
Comp. fit?
Replace Component
Call Instructor
TROUBLE SHOOTING UNIT

Figure 4.--Introduction and "stable-state operation" experimenting procedure

Shlomo Waks
ⓒ 1973

To be practical, make priorities among defined elements.

You must make priorities among defined elements if you have defined many. To make this process easier, you might ask if the defined elements are still related to your producer's doubts and to your evaluation questions.

You could assign high priority to important results, or choose only those defined features of the instructional method which can be manipulated and changed if they are found to be faulty.

Rank high those defined elements on which your evaluator and producer agree. Consensus among producers, subject matter experts, and evaluators is essential to the eventual use of the information gathered about the defined element, for, if there is some disagreement about the definition of a result, for example, a producer may not accept the data collected as a valid indication of the program's success.

* * * * *

Summary

To prepare for a test of your project, you must define those instructional elements embedded in your evaluation questions. You specify the resulting behaviors, course features, audience characteristics, and instructional setting attributes, and the definitions will provide you with the guidelines to choose a test, a sample audience, a portion of the instructional material, and a testing site to be used for the tryout. For that reason, all instructional elements should be defined.

Defining Elements in an Evaluation Question, in Brief

Define...

...goals: to choose or create tests and to estimate the effects of the program in achieving those goals.

...audience characteristics: to choose a sample audience and to account for the effects of the audience on the program.

...instructional setting: to choose a test site and infer what effect the environment may have on the program's results.

...instructional methods and materials: to choose a unit for a tryout.

Define all elements by...

...observing your intended audience working on an early draft or on an alternate but similar instructional method.

Be sure your definitions are...

...specific but flexible.

...comprehensive.

you have considered the full range of results and the makeup of each element by using

--tables.

--mapping sentences.

--flow diagrams.

Choose among elements to consider in a test of the method.

CHAPTER VI

The Tailor's Tape and Assorted Supplies and Tools: The Elements Required for a Test of an Instructional Project

To be most accurate in his work, a tailor must ask a client to try on a new suit. And to adjust a suit so that it fits, a tailor must take precise measurements. If he is commissioned to sew uniforms for a large number of people, to be ready to test his work, a tailor must be ready with measuring tools, a selection of uniforms, a sample of the group for which the uniforms are made, and a place to try the uniform out for its function.

> To be ready for a test of an instructional program, you will need measuring tools, a selection of methods and materials, a sample of the audience, and a place to try the program.

To be most accurate in his work, a project director must ask a student to attempt to learn from the prepared instructional program. To create an instructional project which is effective, a project director too must take precise measurements. But if a project director is commissioned to create an instructional program to instruct a large group of people, to be ready to test his instructional program, a project director must be ready with measuring tools, a selection of instructional materials and methods involved in the program, a sample of the group for which the program was made, and a place to try out the methods and materials.

There are five categories of measuring tools available for constructive evaluation: the review, the progress measure, the criterion measure, the student rating, and the interview.

For a test of a project, you will need a draft of a unit, a sample of your students, and a test site. But measuring tools are the basic elements of a test of a project; the data they provide about the strengths and weaknesses of a project are the essence of a constructive evaluation. From your measuring tools you can find where the strengths and weaknesses are, what they are, perhaps why they are, and what you might do about them.

For constructive evaluation,the types of measuring tools are usually used in the following order:

1. The review: an expert is asked to make a personal judgment about the instructional program.
2. The progress measure: individual students or observers answer questions about the quality of the program continuously while the program is in progress.
3. The criterion measure: individual students are asked to answer questions or to perform in other ways to show they have learned from the program.
4. The student rating: students are asked to express their views of the instructional program on a rating form after the program has been presented.

5. The interview: students and teachers are asked to fully explain the impact of the program as they see it, during or after instruction.

Some of your measuring tools will be simple, some complicated; some of the tests will be subjective and perhaps biased; some will be objective and unbiased. Some of your tools will be administered in a formal standard manner; some will be given informally. In all cases, even in instances of simple, subjective, informal techniques, the evidence secured by your measuring tools must be agreed upon by producers and evaluators as trustworthy for its purpose. In the next few chapters, let us consider each type of measuring tool to determine the purpose for each.

* * * * *

The Elements Required for a Test of an Instructional Project, in Brief

The most complete list of the tryout elements includes...

...five tools
    --the review
    --the progress measure
    --the criterion measure
    --the student rating
    --the interview
...three supplies
    --a selection of methods and materials
    --a sample audience
    --a test site (including staff)

CHAPTER VII

Tool Number One: The Review

Just as a publisher must invite experts to review books, so must a project director call for a review of his instructional methods and materials. While a publisher asks his reviewer to judge the potential market for a book, an instructional developer asks a reviewer to judge the potential effectiveness, efficiency, or acceptability of an instructional method or product. An instructional reviewer is asked to try to anticipate the results of a project and recommend what to do to improve the results.

> A review can result in cost savings and in instructional improvements.

If you call for reviews of your method, even if only for technical problems (quality of speech, clarity of visuals, etc.), you will rid your method of relatively obvious faults so that your tests of the project will reveal more subtle, more important difficulties. You may also short-circuit some costly tests because the information gained from a review can be used to predict the results of expensive field tests. (1)

> A good reviewer is objective, knowledgeable, and practical.

A good reviewer is objective: he reports which of his statements are based on his knowledge and which on his feelings. Sometimes, when there is some difficulty in achieving emotional objectivity, it is better to choose someone to be a reviewer who is not directly associated with your project.

A good reviewer can clearly explain the criteria he uses to make predictions and suggest improvements: few producers would take advice about changing material without a logical explanation. Some quick, sample reviews, written for your staff by various candidates will help indicate a reviewer who can express what he thinks.

A good reviewer is willing to work directly on the project plans and materials. He does not spend his time discussing his abstract views and theories; he applies what he knows directly to the instruction.

CASE

A Trustworthy Review Procedure

In the Communications Research Group at E. I. Dupont de Nemours and Co., researchers spend considerable time developing better ways to evaluate the teaching ability of television commercials.

The first evaluation procedure used in the development of a commercial is a review. A trained researcher scores the commercial's script or storyboard as to its effectiveness by analyzing the content of the commercial according to how well it fits the requirements of a series of construction principles. He reports the strengths and weaknesses of a commercial, his recommendations for improvement, and a Predicted Learning Score. This score is based on the commercial's ability to communicate.

More often than not, one version of a commercial will be submitted to research for review. On the basis of the review the number of versions is cut down, thereby saving considerable cost.

Researchers at Dupont think that review is extremely important. On the basis of an early review tens of thousands of dollars in production costs can be saved. It is done by Dupont staff and does not take long. One of their trained researchers can review a script in one working day.

Ten years of research on hundreds of commercials were done so that now, when a reviewer scores a commercial from

+100 to -100 based on their construction principles, he can be confident that the score will predict field test results. He can also make recommendations which will increase the score. The predicted score equals the field test scores about 70% of the time, and that is good enough to deserve the confidence of producers. (2)

> A reviewer writes a report in which he may state perceptions, predictions, revisions, inferences, principles, policies, or technical remarks.

A reviewer should be asked to write a report which will tell you how to improve your work, but he must present his report early enough during the production schedule so that you will have enough time to make appropriate changes.

A reviewer may include many items in his report. He may...

...state what he perceives. When checking the effectiveness of a television show, for example, to be used to teach reading, the reviewer may report, "That's an interfering stimulus right there."

...predict what is likely to happen. The reviewer may state, "The child's eyes will be directed toward the character on the screen and away from the words."

...suggest what might be changed; he revises. The reviewer may suggest, "Place the character on the right side (from the viewer's point of view) of the words to be read. Have the character orient toward the words and have the words fill the screen; sometimes animate the words."

...state inferences about what factors are likely to contribute to a result. The reviewer might say, for example, "The character is so lively that all the attention will be on him. The words just sit there when they are shown and they change too fast."

...propose general operating rules and design principles. The reviewer may generalize, "Let the word stay on the screen long enough for a normal reader to read each letter separately. Have the character point to and scan the word somehow. Have the word to the left of the character when possible."

...check the method for its fit to policy. The reviewer may say, "That female character is not funny and she's a ridiculous stereotype. We can't let that remain."

...check technical characteristics. A technical reviewer may report that, "That film is too grainy, and if those words are viewed in black and white they'll blend into the background. Also, that fact stated about health statistics is not accurate."

If an administrator is likely to suggest changes, then a policy review should come before an empirical test. But at some times a producer might want a tryout first to secure evidence to present to administrators. (3) For relatively simple, inexpensive instructional methods or products, a policy review would be appropriate after a tryout has been completed.

CASE

Reviewing at an Early Stage Before Student Data is Collected

At the Southwest Regional Laboratory for Educational Research and Development, a federally sponsored agency which produces instructional products, each stage of instructional development is reviewed. The criteria used by reviewers for the decisions they make depend on the stage in the development of the instructional product.

The reviewers are professional staff members and non-laboratory personnel. Competent reviewers for a product might be, for example, subject matter experts, educational measurement specialists, learning scholars, classroom teachers, and curriculum supervisors.

A reviewer, for example, may look at various instructional specifications: 1) a list of prerequisite skills, 2) desired instructional outcomes, 3) a criterion test, 4) a prototype teaching item for each entering skill and desired outcome, and 5) pretest data on pupil performance. He may make a number of different suggestions: 1) a go or no-go decisions to produce the instructional unit based on extent to which learners possess stated outcomes, 2) modifications in sequencing of instructional content, 3) additions or deletions of instructional outcomes and entering skills, 4) changes in criterion items, and 5) collection of pupil data. (4)

The reviewer, in a memorandum, reports what he believes the next course of action should be. When the memorandum is approved, the review at that stage is considered to be complete. The review provides the basis for the next stage in product development.

An example of one format for a review report used at S.W.R.L.

Instructional Specifications Checklist

____1. Specificity of prerequisite skills
____2. Specificity of instructional outcomes
____3. Consistency of stated outcomes with objectives listed in technical plan
____4. Inclusion of all desirable outcomes
____5. Sequencing of instructional outcomes
____6. Completeness and relevance of entering skills
____7. Consistency of test items with stated outcomes and entering skills
____8. Need for additional pupil performance data
____9. Appropriateness of stated criterion levels

Comments and suggested changes:

Recommended action:

> Reviewing an instructional sequence involves compromises among many factors.

A reviewer keeps many factors in mind when analyzing an instructional segment. Often factors conflict with each other and a reviewer may have to make compromises before making a recommendation. A reviewer may reason, for example, that a short, attractive, but non-teaching segment which is a pet project of an administrator, and has cost a great deal of time and money, should be recommended for use if the time is available; that is, if the time is not needed for a segment which leads to one of the important objectives.

Here is a list of factors a reviewer might consider:

Stage of development: has the unit been accepted as a final copy? If it has, the reviewer should be warned that anything he says about change may be ignored or viewed with annoyance.

Cost: how much time and money have been invested? If a tremendous amount of money has been spent, a reviewer's comments about major changes may be ignored.

A producer's personal investment: is the producer willing to make changes or throw out a section based on a review? If he is not, a reviewer should not waste his breath.

Practical flexibility: how many practical changes could be made? A reviewer should avoid suggesting changes which could not possibly be implemented.

Production time left: how much time before the material is needed? A reviewer should only suggest changes which can be made in the time left.

Length and size of section: how many words or minutes? If it's a lengthy section, it may be costly to change.

Curricular relevance: is there a place for the unit in the curriculum; is it redundant; is it unique?

An authority's personal investment: is the unit some administrator's pet?

Social considerations: are there likely to be any side effects which are biased against or toward special interests (women's lib, for example) which are not accounted for?

Educational value: A review must weigh and balance the factors already noted with predicted educational value.

Will it teach well? Does it have any negative side effects? Is it an important objective? Is it attractive and appealing?

CASE

Considering Many Factors in a Review

How does a Dupont researcher do a review? The reviewer reads the script and studies the objective to be accomplished. Then he scores the commercial on a series of construction principles derived from experience and the psychology of communications. Following are some of these important principles. (5)

"INITIAL SIGNAL - This can be defined as what the viewer of the commercial sees and hears in the first one or two seconds of the commercial. It is in this critical timespan that the viewer decides whether to stay with the commercial or go get that beer he's been thinking about during the last half hour. The function of the initial signal is to carry the viewer's attention into the main body of the commercial, and it is on its ability or inability to do this that the initial signal is rated.

DESIGN - This means the kind of development that is used to present the story of the commercial. There are many classes of designs ranging all the way in potential from very strong to very weak. Some examples of these are:

'Problem-Solution' --- in which a problem is presented and solved by the advertiser's product or service. This is a design of high potential.

'Product Display' --- in which little is done in the commercial except to show the product on the screen and describe it in the audio. A typical example of this design is frequently found in fashion commercials. This is a design of moderate strength of wide variability depending upon how well the attributes of the product lend themselves to effective display on television.

'Analogy' --- This design makes its points by some analogous reference to other situations or other materials. It is a design of low strength.

VISUAL DEMONSTRATION - As you will obviously expect, this principle is concerned with what is shown on the TV screen and its relationship to the commercial objectives.

INTELLIGIBILITY AND BELIEVABILITY - This also means just what it says. Does the commercial present its message in a clear and understandable manner? Is there any significant area of disbelief associated with the product, the message, and the commercial?

PERSONAL RELEVANCE - Is the product and the commercial message presented in terms that are relevant to the viewer? Does he really care about it?

SEX-TYPED APPEALS - This principle deals with the commercial content as it relates to the basic sex-oriented drives of men and women. Perhaps you think of these drives as psychological appeals. Examples of strong instinctive drives for women are the presence of children, romantic situations, and situations depicting a women's security. Examples for men include such things as agressive situations, competitive actions, and appeals to mechanical and scientific aptitudes. There are, of course, many more."

A number of scoring points are distributed among the principles. At first there were an equal number of points assigned to each principle, but, as empirical research results came in, relative weights were assigned to the principles, depending on their relative predictive values. Some principles were eliminated, some merged together with others, and some new ones were added.

The score ranges from +100 to -100 and is geared to the scoring system used in field tests of a commercial. When a commercial is shown on the air, researchers call viewers and ask them questions to see if the commercial communicated. Here are some sample scoring points. (6)

"(+100): This person must have learned everything the commercial set out to teach him plus additional information (if present), must have bought the advertised product because of the commercial, and be enthusiastically favorable about the product and the manufacturer.

(+50): Must have learned the main commercial message, displayed an acceptable attitude toward the product, and expressed no unbelievability.

(0): Can prove he saw the commercial but remembers only inconsequential details not associated with the commercial message.

(-20): Can prove he was present during the time the commercial was aired but remembers nothing at all about the commercial.

(-50): A person who left the room during the commercial for a reason that did not demand his presence elsewhere. (A person who left to answer the door or the telephone or because the baby cried, etc., is not scored.) This score is also assigned to a person who is favorably impressed by the commercial and learns everything about the product -- except that he credits it to a competitive brand name.

(-100): This is a person who learned who you are and what your product is and is moved so unfavorably by it that he voices very strong verbal rejection, perhaps with the promise of future rejective actions such as never buying another one of your products or advising his friends not to buy your brand.

> To determine the educational value of a unit, reviewers often use lists of rules, questions, or principles which are based on theoretical or empirically derived instructional principles.

You could ask a reviewer to make predictions based on a set of broad theoretical principles generally supported by research literature. The reviewer could ask himself, for example, if the unit is meaningful: Is the subject matter meaningful for the student? Can the student relate to it personally? Does the material relate to the students' past or present experiences, the students' interests and values, the students' future activities or aspirations, or material to be covered later in the course?

CASE

Reviewing Based on Theory

The staff members at the Far West Regional Laboratory for Educational Research and Development review their instructional products by checking them against rules. They ask, for example, if the learning episode has a clear statement of purpose, ("To see if the child can name colors without seeing an example."); specifies the materials to be used ("Color Lotto Board and one set of colored squares."); and states the procedures to follow ("Say to your child, 'Find a square that is blue.' DO NOT show your child a blue square."). They check to see if a product fits into a sequence of learning activities that proceeds as follows:

(a) free exploration, while the adult observes.
(b) matching.
(c) discrimination.
(d) problem-solving or production. (7)

You could direct a reviewer to make suggestions based on ideas derived from experience and empirical research.

CASE

Reviewing Based on Experience and Past Research

The staff of "The Electric Company" asked the following review questions when viewing scripts, storyboards, and shows, based on their research and observations: (8)

1. Are the words used age-appropriate (Is the verbal humor understandable?)
2. Is the segment short enough to maintain attention?
3. Are references, situations, and words meaningful?
4. Is the educational point obvious?
5. Can the words be seen and heard?
6. Do the words show up in black and white?
7. Do the actors turn toward the words?
8. At a time when the viewer is supposed to read, is there limited action and sustained print to insure that the slow reader has the opportunity to see the words?
9. Are the blends made correctly?
10. Are confusing examples eliminated (e.g., garbage for hard "g" sound)?

11. Are all components (character and action) of the segment consistent within the story?
12. Are there just a few ideas to be taught?
13. Are there repetitions made for the various teaching points?
14. Is the segment socially relevant? (Are setting and character part of the child's normal environment?)

Most reviewers make their comments about educational value in reference to a project's instructional goals, (9) (10) but certain classes of judgment of an instructional system's effectiveness can be made without reference to goals: a program, for example, must be acceptable to a teacher or else it will never be used except to line a closet. A teacher can be asked to review a certain activity to see if it is feasible in his classroom and if the children will be interested in it.

> There are many limitations you must take into account when you use the services of a reviewer.

You want to get an accurate response from a reviewer, but to do so you don't want to change the nature of the instruction by slowing it up or breaking it into artificial sections, for example. If an instructional sequence is changed by stopping it for review, results are likely to be distorted: a break or rest period may boost a reviewer's attention and enjoyment. (11)

A reviewer can become so engrossed in the instruction that he misses some rating points. If he uses a checklist form or a pushbutton to tally his observations, for example, he may lose the flow of the instruction while he is making a check or pushing a button. To compensate for this human error, several reviewers can be asked to analyze units.

You can simplify your thinking by using one reviewer, but you can get a less biased view of your work with several reviewers. When using more than one reviewer, give each one some common questions and some unique ones. Thus, you can compare reviewers' comments and still benefit from their unique abilities.

If a reviewer is trying to take too many factors into account at once, the review may be difficult for him, and the picture of the instruction he shows you may be more accurate, but his perception may be distorted. If a reviewer takes into account only a few factors, the review may be relatively easier for him, but the view of instruction shown to you may be narrowed.

The quality of a review depends on the qualifications of the reviewer.

If a review is of poor quality, the qualifications of the reviewer are suspect, not necessarily the process of review. The title, "authority" or "expert" is usually applied to a reviewer, but the type and degree of authority depends on your purpose. If one is interested in the use of an instructional program to teachers, a teacher is an expert. If one is interested in community reaction, community representatives are experts. If one is interested in student perceptions, students are experts. If predictions about learning are in order, an educational psychologist specializing in the type of learning is an expert. The following is a brief list of possible reviewers and the topics they are qualified to comment on:

1. Students can state their views on the utility and relevance of a method to them.
2. Classroom teachers can state their preferences and personal feelings about the usefulness of a method or product.
3. Production experts can check the technical aspects of a method.
4. Media experts can study the quality of media and its suitability.
5. Subject matter experts can review the quality of content.
6. Experts on learning can compare the characteristics of learning principles to the method.
7. Reading experts can review the comprehensibility and readability of a program.
8. Administrators can provide a policy review.
9. Anyone can review the quantity of content: the number of facts, or the number of physical characteristics.
10. An expert on human development can predict the effects of methods on an audience of a certain age.
11. An expert on sociology or anthropology can predict the effects of methods on a certain type of audience.
12. A curriculum expert can comment on the definition of objectives.
13. Parents and students can review the importance of objectives.
14. A test expert can review the quality of test questions.
15. A panel can provide a broad review.
16. One of the finest and least expensive kinds of review has the producer taking a second look at his own work. (12)

There seem to be three approaches to finding and training good reviewers who are not necessarily classed as subject experts. You can ask a number of potential reviewers to predict results or give opinions, and then see who comes closest to the recorded results. You can go further and give the data to each potential reviewer and see which reviewers use the information to best advantage in their subsequent predictions. The easiest way may be to teach potential reviewers to apply principles employed in validated review forms.

CASE

Using Different Reviewers for Different Purposes

At Michigan State University, Lawrence Alexander, director of the Learning Service, conducted an instructional program which used a combination of peer review and expert review to modify instructional methods of graduate teaching assistants. Each assistant taught his regular class with whatever methods or products he would ordinarily use and his class was videotaped. One camera followed the teacher while the other focused on the class. A technician used a special effects generator to record both images on a split screen.

Once a week each teacher viewed his tapes and selected a short portion which showed what he felt was a problem. A subject matter expert (for example, a math professor for math teachers), and a learning psychologist, and about five of his peers, viewed the selected short portion of tape. An example of one of the tapes might show a five minute explanation of a mathematical principle and a subsequent unsuccessful attempt to get students to apply the principle.

The group discussed what they saw and hypothesized about what might be wrong. Some of the hypotheses might be 1) that while the teacher explained the principle, he did not show how to apply it, 2) that his objective was not clear: it was uncertain whether he wanted students to learn the principle, learn the application of that principle, or learn to apply principles, 3) that students may not have had the proper prerequisites: they did not know the principle, or did not know how to apply principles, or 4) that the explanation was unclear in parts.

They discussed the problem and suggested modifications. The suggested alternatives might include illustrating the principle's application and then asking students to apply it in other ways; making certain that students understand the principle before asking them to apply it, and stating the objective to the students.

The teacher who presented his problem would agree to consider some solutions and select one to try out. He would videotape his attempt and bring it back the following week. (13)

* * * * *

Summary

A review is an excellent assessment technique for use early in project development; its utility depends upon the abilities of the reviewer. If a reviewer can suggest revisions which will improve the efficiency, effectiveness, or acceptability of the program, you will save considerable time and money, and you will be saved extra tryouts which would have led you to the same conclusions.

The Review, in Brief

A good reviewer is...

...objective.

...knowledgeable.

...practical.

A reviewer may report...

...perceptions.

...predictions.

...revisions.

...inferences.

...principles.

...policies.

...technical remarks.

A review involves compromises among factors.

A reviewer may use theory or empirically derived rules.

A reviewer has limitations and the quality of a review depends on the qualifications of the reviewer.

CHAPTER VIII

Tool Number Two: The Progress Measure

Even though most of their activity is not usually recognized, students are very busy while instruction is in progress. A student may be listening, looking, remembering, comparing, making analogies or practicing subordinate skills; or he may be daydreaming, doodling, or talking to a neighbor about his weekend. A student may be reading a vocabulary list of pronouncing words, indicating his progress in learning to read a language. He may be looking at and listening to what is being presented in a language lesson.

A student does not necessarily demonstrate that he is learning simply because he is paying attention, but, because he is attending, one can argue that he stands a good chance of learning. A student's activity during instruction can be represented by many different behaviors; each activity may be measured by several different instruments.

To find out which parts of your instructional method contribute to student learning, you must take measurements during the course of instruction. These are called progress measures, and they are to be contrasted with criterion measures, which are used to reveal what a student has learned at the end of instruction.

You may be able to take a number of progress measures directly by observation.

You may, for example, record a student's attention by observing the amount of time he looks at a page: his restlessness, laughter, verbal activity, and interest. For example, (1) to test for attention and comprehension of printed materials, you may observe and note the pages on which students linger, or you may even attach sensors to the pages as some advertisers do. You may watch students as they perform on classroom practice and laboratory exercises.

You may observe if students approach and stay with instructional materials when given a choice; (2) or check the rate of attrition from one material to another (like Nielsen ratings for T.V.).

> You may take a number of progress checks directly by use of recording equipment.

You may use film, videotape, or time-lapse photography (a movie camera takes one frame at set intervals) to gain a permanent record of instructional events for observation and study at a later time. You may use infrared photography (a still or motion picture camera photographs an audience in a darkened room) to observe audiences viewing films and slide tape presentations. If you have access to the equipment, you can measure eye movement (where a person's eyes focus), respiration, blood pressure, perspiration, and heart rate. You can use mirrors and dual video cameras using the split-screen technique to take pictures of students and instructors simultaneously.

> You may ask a student to participate directly in the progress measure.

You may need several ways for asking a student to participate directly in a progress measure. (3) All will need three components: a signal to a student to respond, a simple way for a student to respond, and a way of recording the student's response. For example, you can ask a student to write down his answers to classroom exercises, or you can ask a student to answer a continuous question at timed intervals: are you learning and are you enjoying the instruction? A student can record his answer on a form like this:

TABLE: Form for student response during a lesson

Check the appropriate category when the number is flashed.		Learning	Not Learning	Confused
	1.	☐	☐	☐
	2.	☐	☐	☐
	3.	☐	☐	☐
	4.	☐	☐	☐
	5.	☐	☐	☐

Another simple technique for continuous recording of student responses during instruction is to have students, at a signal, mark a space on a chosen scale; for example, "Check one of these: like - indifferent - dislike, learning - not learning - confused." (4) (5) (6) Students may be signalled to write down their ratings by a slide projector flashing a number on a screen.

CASE

Asking Students to Participate in Recording a Progress Measure

At Children's Television Workshop, Keith Mielke, at the time a Spencer Fellow, suggested many tests for assessing comprehension of information as it is presented. (7)

He recommends asking a student a question before the instructional segment. A student must keep the question in mind during instruction and answer it as soon as he knows; his response should be timed precisely. It has been found that asking a student a question before instruction does direct the student to look for certain points, and it tells a producer when the learning is taking place, but the measure should be considered an overestimate of the teaching ability of the segment.

He suggests asking a student a question as instruction progresses. In this case answers should be given individually.

Mielke says that when given only the audio stimulus or the visual stimulus of an audio-visual presentation, you can ask the student to tell what is on the missing portion. Delete information in either audio or visual and ask a student to supply what is missing. This way you can locate learning in the segment precisely. Finally, Mielke describes a test in which you stop the presentation and ask a student what led up to that point and what is likely to come next.

Here is an example report on a comprehensibility study done by "The Electric Company" research staff.

MEMORANDUM
CHILDREN'S TELEVISION WORKSHOP

DATE: Mar. 19, 1973

TO: John Boni, Sara Compton, Tom Dunsmuir, Thad Mumford, Jeremy Stevens, Jim Thurman, [writers], Tom Whedon, [Head writer] & Andy Ferguson [Producer]

CC:

FROM: Research

SUBJECT: Comprehension Study

Some time ago we did a short study to find out how comprehensible a typical (as opposed to experimental) show is to children who do not watch the "Electric Company" regularly, if at all. We showed the show (#203) to twelve children, stopping the tape after different bits, and simply asked them questions such as "What happened?" "Who is that?" "Why did they do that?" etc. If the child did not mention an important aspect of the bit spontaneously, we asked him about it specifically. The following is a summary of the results:

Balloon Blending:

Almost all the children knew that balloons were being popped, but only two of them mentioned that there were any words under them. When asked about the words specifically, some remembered PIG, some POP, one child also remembered PET, but almost half the children had failed to notice the words at all.

P-Pickin Song:

We asked the children which letter all the words had in common (after explaining to them that they had a letter in common, which they usually have a hard time realizing) and almost all of them knew that it was the letter P. One child thought it was B, but that might even have been a pronunciation difficulty.

Prim/Proper:

Only one child knew what "prim and proper" means.

Archie Bunker Cameo:

None of the children knew who Bunker was, and none of them knew what he had said ("Stifle yourself").

Lilly Tomlin Cameo:

None of the children knew who she was, and only one child knew that she had said "That's the truth."

Pain:

All the children told the story accurately, and all of them knew that the boy had a pain. They also all recognized the doctor and nurse, and all but one knew that the patient was a football player.

"ay" Machine Animation:

Only one child knew which letter combination went into and came out of the machine.

Vi's Diner:

This was the super-supper bit. Half the children said that "there was something wrong with a word", and the other half knew what the word was. Half of them knew that the man's job was "word-repairman", and almost all of them knew that Vi was "a person who works around food."

Joe Namath:

Half the children had no idea who he was, and the other half thought he was either a football man or a baseball man. Only two knew that he had said the word PASS.

Cosby as Prince:

All the children knew that Cosby had forgotten his pants, and most of them realized that he had a note, but only half of them knew why his wife left him the note, and they thought it was to remind him to put his pants on. That is, the children did not realize that the joke was supposed to be about forgetfulness to an absurd extent.

Letterman:

All the children knew that an octopus figured in the bit, but only half of them said that the villain turned the bus into an octopus, and that the hero turned it into something else. None of them could say what he turned into. Only half of them knew Letterman's name. Half of them knew that he changes things into other things, some of them knew that he is the good guy, and one of them knew that he changes letters. None of the children knew Spellbinder's name, but most of them knew that he changes things into other things.

CASE 2

Asking Students to Participate in Recording a Progress Measure

In a course being developed in the psychology of learning, a professor asked students to record their answer to classroom exercises, consisting of a principle or two, which were given after each time interval used in class. After teaching each of the principles of behaviorism, he gave a short exercise to find out if the students could choose the attributes of the principles. Questions were put in this form, "The distinguishing characteristic of positive reinforcement is..." On one part of the exercise the student had to choose from attributes given, and on the other he had to supply the answer.

On the next short exercise, after a bit more instruction, students were asked to choose or supply an example of the principle taught. The questions were, "Choose from these examples the one which illustrates positive reinforcement," and "State an example of positive reinforcement."

After additional instruction the professor asked his students to state how they would apply the principle by choosing correct applications, how they would apply a given principle in a situation, and what they would do in a situation (with no principle given). Questions were phrased in this way: "In one of the following applications, the teacher is using positive reinforcement correctly to encourage reading. Which one is it?" "A teacher found that his class would not do complete homework assignments. How would you use positive reinforcement to get them to hand in complete work?" and "A teacher found that students were not coming to their reading groups as quickly as they might. What would you do to get them to move more quickly?"

Finally, after some more instruction, students were asked to apply any of the principles they had learned to a real-life problem they could find with a neighbor's child, their own child, or in a classroom to which they had access.

Even if the professor could not observe the students as they did the exercises he had a permanent record of the students' progress and the success of each instructional segment. He found that a number of students decided to punish a child in a given case unnecessarily when they could have used positive reinforcement only. He traced back to see if they could distinguish the attributes of the principle and identify examples; he found that they could. Then the professor checked to see if his students had any trouble choosing correct applications for a given situation. He found that many of those students who did not do well in the case situations did not know the situations in which to apply a principle. At this point he reviewed the instructional segment which preceded that test and looked for possible contributing factors.

You may ask a student to respond using recording equipment -- by pressing a button or tapping a foot-pedal as the Health Show researcher used. You can prepare visual material such as slides and film so that the slide brightness fades unless a foot-pedal is pressed; if a student lets it fade to minimal brightness the next visual segment appears. Since attention level is defined as the number of presses made by the subject during the first 6.5 seconds of exposure, (8) you may, for example, infer appeal when you ask a student to maintain the brightness and clarity of a slide by pressing a foot-pedal. (9) (10) Other techniques have been suggested: an audio switch, a dial to indicate a response, (11) and a dial to reduce noise.

CASE

A Technical Procedure Used as a Progress Measure

If you want to know precisely where a single subject is looking, you can use eye movement patterns. Experimental psychologists have instruments which include a helmet-mounted corneal reflection system which transmits the eye movement data to film or tape. The finished product is a film of the visual presentation with a white dot superimposed on the spot where a student was looking at a given second. An excerpt from a preliminary eye movement research on "The Electric Company" follows: (12)

"1. Show #206 (extended duration of print):

Poor readers do get through the scanning.
Kids reading near grade level get bored pretty fast.
Good eye movements for kids at all levels.

Improvement of scanning print from one bit to the next when the same curriculum piece is presented is probably exponential. (i.e., if first scan takes 4 seconds, second will take 2 seconds, third will take 1 second, etc.) Once kid has got it, he reads it again and again.

See Sam Calypso: While actor sits still, kids scan print. As soon as he moves, all eye movements centrate [center] on him. Suggests that it may be good idea to go through whole sequence static, then activate.

Time on screen for print optimal in this piece.

2. Not Safe for Swimming:

Did not work well. Very scattered eye movements. We may be modeling poor reading. More centration on errors than on correct sequence of print. Least effective piece tested from point of view of centration on print.

3. Clowns:

Best eye movement when clown stood still. When he mugs and gestures, eye moves away from print. May not be bad for short items, like blends, as in this piece.

4. Silhouettes vs Faces Blending:

Silhouettes much more effective (p .01). Excellent technique, however, both ways. Kids like very much. Probably our best blending technique. Interesting additional point: male face more interesting; a lot of scanning of beard.

> You should observe student activity during a program at frequent intervals, so that if your instructional procedures are faulty, you will be able to spot the resulting inadequate student performance as near to the procedural fault as possible.

Because the important requirement of progress measures is to record observations in relation to critical points in the instruction, an evaluator and a producer must decide at what times during the instruction to record their observations, how to index students' responses, how to synchronize the instruction with the record of students' responses, and how to score student behavior.

CASE

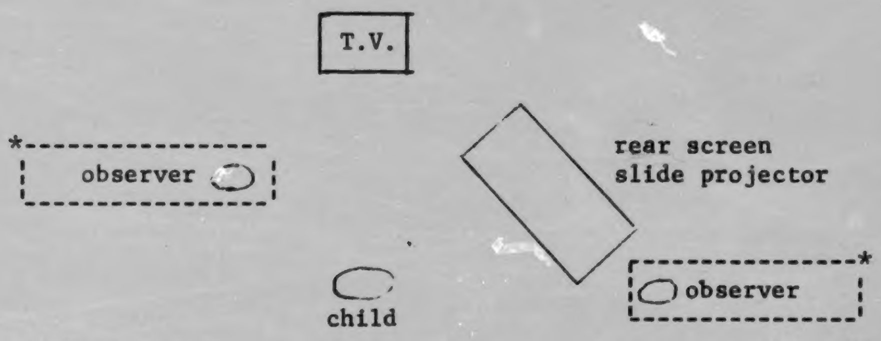
Observing Student Activity at Frequent Intervals

Early in the development of "Sesame Street" the producers and researchers realized that in order to help 2-5 year olds learn from television, they would have to capture their attention. Thus, one of their first evaluation questions was asked: "Can we hold the attention of young children?" In this case "attention" was defined as looking at the source of instruction, the T.V. screen; attention to the audio portion was not included. Visual attention was considered especially important for "Sesame Street" because the content of the instruction was primarily word-symbol correspondence: letters, numbers, sight words, labels for processes, the concepts "alike" and "different".

To define attention in a fashion that could be useful for creating a measure, the setting had to be taken into account. The setting included a child sitting in a room where other children and adults might speak, move around or otherwise distract the child from viewing. To represent this condition the amount of visual distraction was standardized. Thus, the definition of "attention" was looking at the television screen while a visual distraction was also available. (13)

To measure this definition of attention researchers at the Children's Television Workshop use a test which they call "the distractor measure." The distractor measure is a progress measure because it assesses student behavior during the instruction. As it is used at C.T.W., a rear-screen slide projector is placed adjacent to a television screen, and at a 45-degree angle. Slides, randomly placed in a carousel, change every seven and one-half seconds. An observer records the seven and a half-second intervals during which the child's eyes are looking at the television.

Figure - Diagram of placement for distractor, T.V., observer and child for distractor technique.



*alternative positions

(14) If the child's eyes stay on the set for seven and one-half seconds, a 3 is assigned the interval. If his eyes stay on the set more than half the time, a 2 is assigned. If his eyes stay on the screen less than half the time, a 1 is assigned. If during the interval his eyes are never on the screen, a zero is assigned.

To do a distractor study on a television show, an evaluator must be aware of some techniques:

The evaluator must set beginning times and check points throughout the show before the study. For example, two observers might agree that the first 7½-second interval begins when the show number is on the screen and that at the beginning of a Bert and Ernie sequence is the beginning of observation 20 (the 20th 7½-second interval). When Bert and Ernie come on, an evaluator can check himself to see if he has been keeping up. If he has missed recording a score for a 7½-second interval and finds his last score recorded for interval 18, he moves to record a score for interval 20 and continues. If there were no checkpoints, the summarized scores from many observers would be full of errors. Indeed, if any interval is missed, each interval after that would be mis-scored. For example, if the third interval was missed, the next interval would be scored as the third when it should be the fourth, etc.

The click of the carousel slide projector tells when the new interval begins. It is useful to have a simple counter wired to the projector to let the observer know at a glance what interval he's scoring.

There are several recording methods which can be used. The observer can look down to write, press down on a continuous recorder when the child is looking and let up when he's not, press one of four score buttons (0,1,2,3), or whisper or tap into a tape recorder microphone. If the sound is turned up loudly enough he will be able to coordinate the exact spot in the show with his student's observation. The first method, and perhaps the second and third methods, may have the disadvantage of drawing the observer's eyes away from the child. The last method will not draw the attention of the observer away from the child and will not be so audible as to distract the child. In some cases, when eyes are drawn away, observers watch for an interval and then record for an interval. They feel that they are trading accurate recording every other interval for recording all intervals with a larger chance for error and some loss of observer time.

The observer should sit out of the line of sight of the child, but close enough to see where the child's eyes are looking.

Older children may be capable of attending to two things at once. They may be able to take their eyes off the screen, look at the slides, and still get the message. An observer can double-check his work by recording the child's behavior by videotape or super eight film running at regular pace or at timed intervals.

An observer may double-check his measure of attention to find out if a child paid enough attention to get the visual message. This measure combines attention and memory, but can be useful in interpreting the distractor data: play back the audio portion and ask the child to describe the visual.

When instructing children before a distractor study, be sure to tell him that it is perfectly all right to watch the slides if he wants to. If an observer says nothing about watching the slides, the children will watch the show because they were told to, not because it was more appealing than the slides.

A carousel of 80 slides is usually used. With larger trays of slides available, an observer may decide to use more than 80: more slides may prove to be a better distractor because the slides continue to be relatively novel. Researchers usually buy assorted sets of slides and mix them, and if they plan to reuse the slides many times, choose plastic framed slides rather than paper ones. Other distractors (magazines, toys) have been used but not with large numbers of children, but other evaluators may find these valuable.

The number of children observed at once depends on the number of children in the natural setting. If project producers expect one child to learn alone from their materials, use one child. If the method is to be used in school, use a small group. Each additional child introduced is an additional distraction. When groups watch, the overall average of attention for each segment drops, but an attention-getting segment still scores relatively high and a low segment, relatively low.

The time interval used to observe was chosen because it gave the viewer time to react, and react just long enough for his behavior to be classified. If an observer were to wait a longer time, a great deal of information about a person's looking would be lost. In addition, many "Sesame Street" segments are only seconds long. If the interval were longer, a whole segment might get only one observation. To find out what happens to attention within a segment then, a short interval of observation is needed.

"The Electric Company" researchers in some informal research, compared the resulting averages of attention taken from observations of children at one-second, five-second, and ten-second intervals. These results were also compared to continuous recording data made by pushbutton techniques, and a casual observer's recordings. The usual results for five-second intervals were similar to the one-second intervals, while results from ten-second intervals were quite different. Time sampling results were more like the continuous recording results than were the casual observer's recording. A time sampling procedure using a 5-second interval may be an accurate and efficient distractor recording procedure.

Setting up, recording, and scoring are the collection procedures in the distractor method, but the technique consists of more. The distraction data is summarized and profiled as follows. The data for several children is added by time interval. The base of the graph consists of the time interval observations, by number. The vertical axis consists of scores from 0 to the potential 100% appeal score. If 10 children were observed, a 100% score during an interval would be 10 scores of 3, or 30, representing continual attention. Across the top of the graph, notes about the segments may be made: "animation number 3: barnyard." At sharp peaks or troughs in the line drawn between the summarized scores, additional notes may be stated: "music started here".

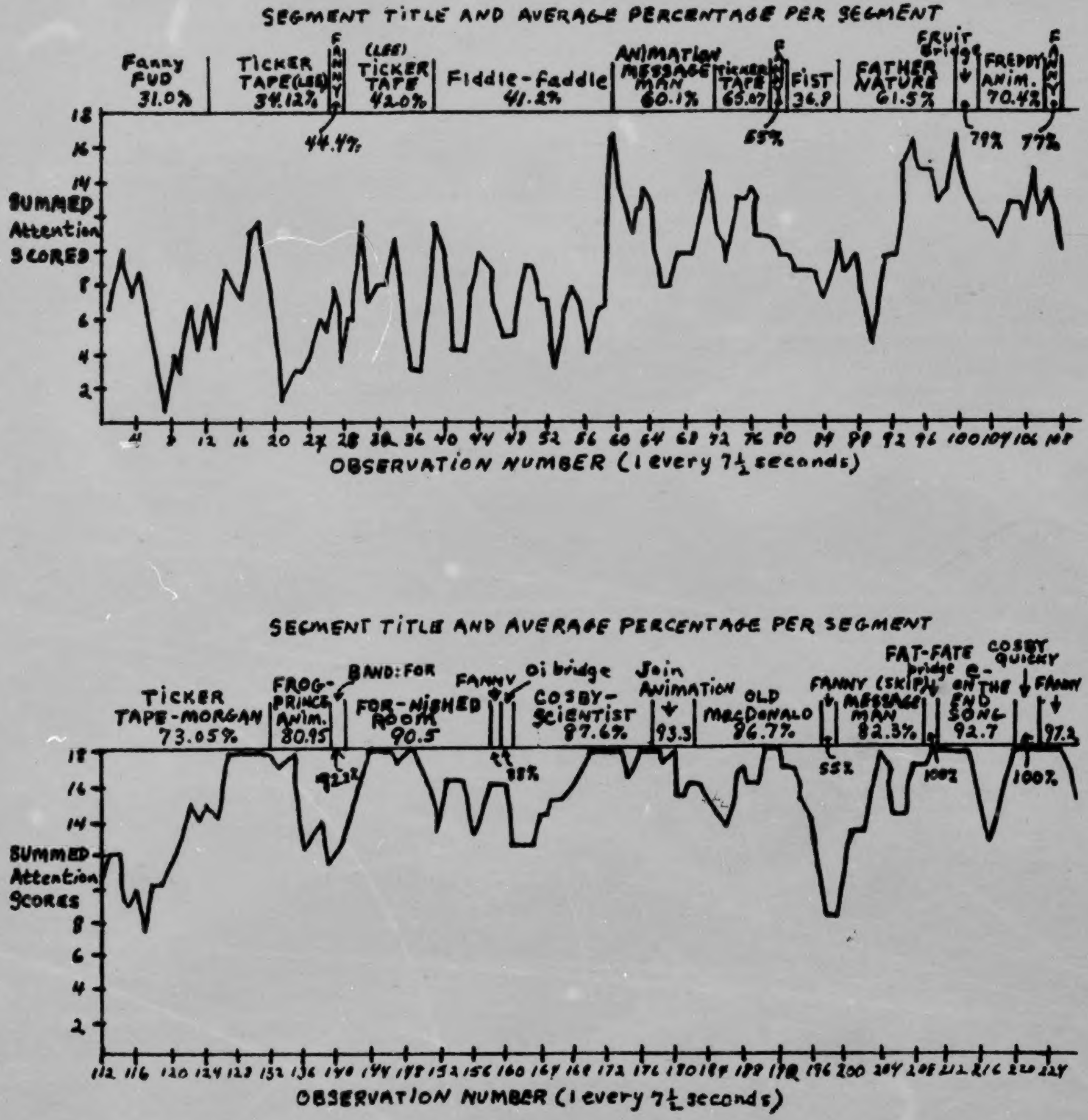
Then an evaluator computes an average attention level for each segment and for the entire group of viewers over the course of the program. The summarized scores for each segment are averaged to get a "segment average", and then these are averaged. This result is the show's report card. A show gets

> an overall grade, 76%, for example, and each segment gets an average score. An evaluator may include a number to indicate his degree of confidence in the results.
>
> The scores are meaningful to writers and producers. They feel that a 70 is fair, an 80 good, and a 90 a fine show. But these scores can remain meaningful only if the test procedures are consistent from test to test. The same number of subjects, the same distractions, and the same scoring procedures should be used from test to test.
>
> The purpose of the distractor is to help develop hunches about a show: from distractor measurements have come some specific and some general hunches. Specific hunches relate only to individual segments: "That bit dies after they start the dialogue." Some general ideas are formed, too: "Attention is high for segments with animation." The general comments help in the design and redesign of instruction.
>
> Every test has its advantages and disadvantages. A distractor measure is useful and its results do not require elegant interpretations. The profiles called distractographs provide a brief summarized view of many responses to a program. In addition, individual segments can be studied from moment to moment. To a limited degree programs can be compared to each other if students and conditions are the same or randomly assigned from the same population.

Progress measures like the one described are useful. They provide immediate feedback to producers about the attention levels of an audience each minute. Producers can use this evidence to trace the sources of strengths and weaknesses in a show.

There is some controversy around the validity of progress measures. Many educators feel that these cannot be taken too seriously until more precise methodological research has been done. But each evaluator should judge for himself and consider some of the evidence. For example, consider the progress measures indicating "Physiological Arousal During Instruction." Some research has shown that high arousal is associated with remembering and low arousal with forgetting; other research has shown that a student may like instruction and learn little, or that a

Figure: DISTRACTOGRAPH FOR THE ELECTRIC COMPANY SHOW 133



SHOW AVERAGE = 67.74

SIX SECOND GRADERS FROM A NEW YORK CITY SCHOOL - SAMPLE student may find a subject interesting and learn nothing. His attention can be high and his learning low; his visual attention may be low but he can still learn by listening. But he must pay some sort of attention for learning to occur at all. When a student says he is learning, his scores on a test will reflect a higher degree of learning than if he reports no learning.

In general, a progress measure is a valuable tool for a constructive evaluation. With the evidence gathered from progress measures and scores, an evaluator and producer can find the sources of the success or failure of an instructional program.

* * * * *

The Progress Measure, in Brief

Use progress measures by...

...observing directly.

...recording behavior with mechanical equipment.

...asking students to participate.

...observing at frequent intervals.

CHAPTER IX

Tool Number Three: The Criterion Test

An instructional program is a success when students learn; to gauge the success of a program is to find out what and how much students have learned. Although students may learn many things from any lesson, a teacher is primarily interested in observing student performance related to his instructional goal; criterion measures are needed for this task.

> Criterion measures reflect objectives and the possibility of unforeseen results.

Criterion measures must cover all the defined objectives and must also include other behaviors in order to explore the possibility of both positive and negative unforeseen results.

One possible criticism of criterion tests is that teachers end up teaching for the test. If the test really tests important things like the diagnosis of disease, then there is absolutely nothing wrong with a test that mirrors the objective and the teaching. If, however, students are taught concepts, test items should require the student to identify examples of the concept not used in the program. When test items deal with principles that require students to make predictions or explanations, the items should include situations not covered in the instructional program. In sum, if the objective and the teaching are really important, then there is nothing wrong with a test that mirrors them.

There are six major steps necessary to create a criterion test.

First, (1) you describe the learning desired. Next, based on the behavior desired, (2) you choose a test format. You may, for example, require a choice of answers or require the production of an answer. Then (3) you write several test items to measure each objective. You must have a sufficient number of quality test items to permit the teacher to interpret a student's test performance as mastery of the objective.

If a test question asks for knowledge or performance which anyone might know, it's not worth asking. The correct answer to a simple item or a single item will not convince most teachers that a student has mastered the objective: a student might appear to have mastered an objective, when, in actuality, he guessed.

Once items are written, (4) check to be sure all the topics and behaviors are covered. Check to see if you have asked questions calling for all the behaviors or topics taught.

When the list of items is complete, (5) you form the criterion test by assembling items in groups according to objectives, easiest first. Finally, (6) you set a cut off point of acceptable performance on the test -- 80%, for example.

Criterion measures should be varied.

Most project directors create criterion measures which are typical school tests, but they need not be: criterion measures may take many forms. Various forms of assessment can be used; for example, to measure the extent that a program succeeds in achieving the goal of a certain student attitude, an evaluator could ask the student about his behavior; he could ask which of these two instructional products the student would read, listen to, or look at, or what the student is likely to do at a certain time. Using another method, an evaluator could observe student's behavior as the student works. He might check the amount of time spent outside class on the subject, the student's comments, facial expressions, and body movements. He could record the student's reactions with instruments -- an eye-movement camera or a polygraph. He could provide two presentations and ask subjects to choose one. (1) He could ask the student to choose descriptions of the subject in question or rate it on a 10-point scale. All of these may be considered acceptable criterion measures for a change of attitude if they fit the defined objective.

CASE

Creating a Criterion Measure

The staff at the Southwest Regional Laboratory for Educational Research and Development developed an instructional concepts program to teach 86 concepts to kindergarten children. (2) Research staff members at the laboratory reviewed a number of first grade curriculum guides to compile a list of the concepts that a kindergarten child should know. They found many concepts embedded in the teachers' instructions. For example, the curriculum guide might suggest that the teacher tell the children to look at the top of the next page. "Top," "next," and "page" must be understood before the children can follow the instruction. The researchers revised the original list based on the advice of teachers and curriculum specialists. The final list contained 86 concepts grouped into seven classes: color, size, shape, position, amount, time, and equivalence. The goal was to have children learn to comprehend these concepts when they were presented orally.

The first version of the instructional concepts program included 32 lessons; each lesson consisted of a story and posters illustrating a concept. Optional activities (games, flashcards, and practice exercises) were available.

Tests were constructed to assess the success of the program. One criterion test measured the ability of the children to identify concepts. Identifying a concept was defined as pointing to a picture illustration of the concept name, when shown with two other examples.

Because each child could not be asked 86 questions, some sampling of concepts was necessary. Five concepts were randomly chosen from each major list of concepts. One item represented each concept selected.

It should be noted that the researchers could have used other sampling techniques. They could have used all concepts and written more than one item for each concept. Not all children would have had to take all items, but all parts of the program could have been tested out on a number of children.

Examiners asked children to point to an illustration of a concept when it was presented with two non-examples. For example, children were asked, "Point to the green bird." "Point to the bowl with the most ice cream," and "Point to the monkey at the beginning of the line."

Eight classes of children were tested before and after the program. When the scores were corrected statistically for guessing, the results showed a move in average percent correct from 49% before the program to 70% after the program.

The scores also revealed particular strengths and weaknesses related to different concept goals. Children learned most from the program about shape and position concepts, a gain of 30 and 28 percent respectively. Children learned least about size concepts (11% gain). By the end of the first version of the program children knew all their colors (96% of concepts were identified), but knew relatively little about equivalence (54% were identified).

The results gained from the criterion test showed what the program had taught and what it had failed to teach. The results did not show why the strengths and weaknesses appeared: these are the limitations of criterion measures. Other measurement techniques are necessary to find answers to those questions.

Summary

Although criterion measures are only one of the tests used in a constructive evaluation, the criterion test provides the most convincing evidence as to which parts of the program work and which do not. (3) Performance on criterion measures show that a problem exists and what the problem is, but not where or why. But other types of tests such as progress tests can provide unique contributions to program diagnosis.

The Criterion Test, in Brief

Criterion tests reflect objectives and the possibility of unforeseen results.

There are six major steps to create one:

1. Describe the results.
2. Choose a test format.
3. Write items.
4. Check for comprehensiveness.
5. Form the test.
6. Set a cut-off for acceptable performance.

Criterion tests should be varied.

CHAPTER X

Tool Number Four: The Rating Form

There are some insights into the strengths and weaknesses of an instructional program which can be secured only by asking students to indicate their thoughts and feelings. Their perceptions and opinions can be stated on a rating form or questionnaire.

> A rating form is an efficient way of getting many useful ideas from many people at one time.

A group of students can be asked many things at once:

1. What were the course goals?
2. How well were objectives identified?
3. Was the program effective? How did it influence your...
   a) choice of major
   b) electives
   c) decision to study further
   d) job decision
   e) preparation for work
4. Were the objectives reached?
5. How did methods contribute to learning?
6. Was all content covered?
7. Was all content appropriate?
8. Was it enjoyable?
9. Was the instructor enthusiastic when presenting course material?
10. Did the instructor seem to be interested in teaching?
11. Did the instructor use examples or personal experiences which helped to get points across in class?
12. Did the instructor seem to be concerned with whether or not the students learned the material?
13. Was the instructor friendly and relaxed in front of the class?
14. Did you feel this course challenged you intellectually?
15. Were you generally attentive in class?
16. Did the instructor encourage students to express opinions?
17. Did you have ample opportunity to ask questions?
18. Did the instructor appear receptive to new ideas?
19. Did the instructor attempt to cover too much material?

20. Did the instructor lecture above your level of comprehension?
21. Could you see how the concepts in this course were interrelated?
22. Were the class lectures made for easy note-taking?
23. Did you know where the course was heading most of the time?
24. Was the grading system adequately explained?
25. Were the answers to exam questions adequately explained after the exam was given?
26. Were course objectives reflected in the exams?
27. Could you see how the course material could be applied to your personal problems?
28. Could you see how the course material is pertinent to your major field of interest?
29. Did the instructor make you aware of current problems in the field?

Here are some typical responses to a very simple end-of-the-class questionnaire:

1. What did you like best about this class?

   Sample student responses: "Clearly stated objectives."
   "Informality of the class."
   "Opportunity to ask 'stupid' questions."
   "The examples given."
   "The lectures are getting more relevant, or at least I understand them better."
   "A chance to see alternative ways of solving the problem."

2. What did you like least?

   Sample student responses: "Please go slower on explanations."
   "Information was not clearly explained in proper order."
   "Too much technical material at once."
   "Some people monopolize the discussion."
   "The room was too warm."
   "Too much jargon without explanation."
   "The tension of waiting for a turn to report; of finding out what I did wrong and have to redo."

3. What did you accomplish?

   Sample student responses: "I made up my 'head' about my project."
   "Verified that I was on the right track with my project."
   "I learned to be more specific in my approach."

4. What changes in class procedure would you suggest?

Sample student responses:	"Confusion in class discussion could be cleared up by explaining rules." "Give more examples." "Arrange time for students who are bogged down with problems to come into your office for help." "Work in smaller groups with the instructor." "More time to work independently."

5. What specific questions do you want answered?

Sample student responses:	"What is a ____________________?" "Do we have to revise old material as we get new ideas or make new decisions?" "Is it possible to have class on a different night?"

You may use the summarized results of rating forms as the basis for group or individual discussions about the program's features. You may then direct discussions to elicit hypotheses about the reasons for program strengths and weaknesses and perhaps ask students to suggest ways to improve.

> Rating forms should be integrated into the usual course of the program, should contain specific content and criteria, and be formed to show what changes are to be made.

The act of rating should not interfere with normal reaction to the program. It is possible that when students are placed in the role of raters, they attend, enjoy, and comprehend the subject matter in a much different way than they would if they attended class merely "to learn." (1)

The usual student opinion is marginally useful for the evaluation of instruction: students score generously, are not frank, and report indirectly. Therefore, the content and criteria of the rating form should be as specific as possible. Unless criteria for each rating are spelled out, student raters are likely to have difficulty with their evaluations because their impressions are likely to be determined by the entire instructional program, rather than individual segments or aspects.

The rating forms should be constructed to imply that corrective action will be taken. Patricia O'Connor of the School of Dentistry at University of Michigan, designed evaluation forms to provide clear implications of changes to be made. The test included items about the appropriateness of objectives, their attainment, and testing. Students were asked to describe critical incidents where teachers did something helpful or detrimental. The results speak for themselves:

> "...in a practice management course, students rated the relevance of each project to dental practice and stated information and skills they wished to acquire. Most projects were rated low and new skills and information were identified. The instructor eliminated projects, scheduled lecturers from other disciplines and is developing criterion tests and instructional materials simulating decision making in private practice." (2)

> "...In a course in dental hygiene, critical incident data and responses to other questions revealed problems in consistency among instructors in recommended procedures and evaluation. The course director developed videotapes demonstrating procedures and supplied faculty and students with statements of objectives and assessment instruments. The following year, statistically significant (t test) improvement was shown in questions concerning staff preparation, flexibility, knowledge and enthusiasm, but not in attributes unrelated to changes introduced." (3)

CASE

Using a Rating Form for Project Improvement

The following case is an excerpt from a doctoral dissertation by Allan Abedor at Michigan State University. In his thesis Abedor investigated an approach to constructive evaluation which included the use of a rating form. The purpose of the form was to acquire quick, summarized information about general student reactions to a lesson. The results were used as the basis of a group discussion.

Abedor was working with a few college teachers who had prepared SLATES. SLATES is an acronym for Structured Learning and Training Environments. A SLATE consists of varied materials, texts, slides, tapes, films, or manipulable materials.

The materials are presented to students in individualized self-administered packages, each containing several lessons which help the student achieve some specified objectives; at Michigan State University, students have learned soil science, observation skills, teaching skills, music, cattle identification, and nursing skills by SLATES.

After reviewing a professor's SLATE for technical flaws, Abedor administered the program to individuals when possible, or to a small group when necessary. For example, he gathered 10 students together to view a SLATE consisting of a slide and tape presentation on cattle breeding. The students were asked to work during the presentation as they would in class. Abedor observed and noted questions and signs of inattention and discomfort during the SLATE. When the program was over, the course professor (the producer of the SLATE) asked students to take a short criterion test and a rating form.

The rating form was constructed by Abedor for the specific purpose of finding strengths and weaknesses in SLATE programs. He asked questions on rating form items which related to a number of important factors: ability of the SLATE to communicate, ability of the SLATE to teach, ease of use and ability to influence attitudes. Study the questionnaire and then look at the way Abedor classified his rating form items. (4)

STUDENT REACTIONNAIRE

NAME________________________________DATE______________

LESSON TITLE____

Please be frank and honest in answering the following questions. Remember, you are our prime source of information regarding what needs to be revised.

KEY: 1 means you strongly agree; 2 means you agree; 3 means you are uncertain; 4 means you disagree; and 5 means you strongly disagree.

1. I had sufficient prerequisites to prepare me for this lesson. ____ ____ ____ ____ ____
   1 2 3 4 5

2. I was often unsure of what, exactly I was supposed to be learning. ____ ____ ____ ____ ____
   1 2 3 4 5

3. After completing the lesson, I felt that what I learned was either directly applicable to my major interest, or provided important background concepts to me. ____ ____ ____ ____ ____
   1 2 3 4 5

4. Manipulating the equipment, or equipment breakdowns, often distracted my attention. ____ ____ ____ ____ ____
   1 2 3 4 5

5. Listening to the tapes and watching the slides became tedious or boring. ____ ____ ____ ____ ____
   1 2 3 4 5

6. This lesson was very well organized. The concepts were highly related to each other. ____ ____ ____ ____ ____
   1 2 3 4 5

7. A professional speaker (announcer) should be used to make the tapes. ____ ____ ____ ____ ____
   1 2 3 4 5

*8. The audio tape moved too fast for me: there was too much information. ____ ____ ____ ____ ____
   1 2 3 4 5

*Some of these questions could have been phrased more precisely--many have two questions in one.

9. There was too much redundancy. I was bored by the repetition of ideas.
1 2 3 4 5

10. There was a lot or irrelevant information in this lesson.
1 2 3 4 5

11. The workbook was excellently designed. I could easily follow the instructions and perform the exercises.
1 2 3 4 5

12. Frequent reference to and use of the workbook was distracting.
1 2 3 4 5

13. Often the tape and slides seemed unrelated to each other.
1 2 3 4 5

14. This lesson had very serious gaps and lacked internal continuity.
1 2 3 4 5

15. The examples used to illustrate main points were excellent.
1 2 3 4 5

16. The vocabulary used contained many unfamiliar words. I often did not understand what was going on.
1 2 3 4 5

17. The pre-test and final exam questions did a good job of testing my knowledge of the main points in the lesson.
1 2 3 4 5

18. The questions during the lesson gave me valuable feedback on how I was doing.
1 2 3 4 5

19. Many of the things I was asked to do, or questions I was asked during the lesson, seemed like needless busy work.
1 2 3 4 5

20. At the end of the lesson I was still uncertain about a lot of things and had to guess on many of the final exam questions.

____ ____ ____ ____ ____
1 2 3 4 5

21. I believe I learned a lot, considering the time spent on this lesson.

____ ____ ____ ____ ____
1 2 3 4 5

22. I would recommend extensive modifications to the lesson before using it with other students.

____ ____ ____ ____ ____
1 2 3 4 5

23. For you, what was the most difficult part of the lesson?____

____

24. What was the easiest part of the lesson?________________

____

25. What were the three worst things about this lesson?________

____

____

____

26. I understood most of the concepts and vocabulary immediately after completing the lesson.

____ ____ ____ ____ ____
1 2 3 4 5

27. I think this whole procedure of trying out new materials with students is a waste of time.

____ ____ ____ ____ ____
1 2 3 4 5

28. I would prefer a textbook or lecture version of this lesson rather than the slide/tape/workbook version.

____ ____ ____ ____ ____
1 2 3 4 5

29. I often needed to go back over a portion of the lesson to fully understand it.

____ ____ ____ ____ ____
1 2 3 4 5

30. After completing the lesson, I was more interested in and/or favorably impressed with the general subject matter than I was before the lesson. ____ ____ ____ ____ ____
1 2 3 4 5

31. Please write below any comments, suggestions, or changes which you believe will improve this lesson. Thank you.

The Relations among Questions in Abedor's Reactionnaire

1. SLATE strengths and weaknesses resulting from communication/message design factors:

Factor	Item Number
a. Rate of presentation	8
b. Redundancy	9
c. Interest and attention	5
d. Clarity of instruction and examples	11,13,15
e. Vocabulary level	16
f. Audio and video quality	7

2. SLATE strengths and weaknesses resulting from learning or task factors:

a. Prerequisites	1
b. Objectives	2
c. Motivation	3
d. Organization and sequence	6, 14
e. Evaluation and feedback	17,18
f. Type of response and frequency	12,19
g. Relevancy of information	10

3. SLATE strengths and weaknesses resulting from management/technical factors:

a. Equipment manipulation	4
b. SLATE methodology	28
c. Tryout procedures	27
d. Degree of revision needed	22

4. Perceived learning and attitudes resulting from the lesson:

a. Attitude towards subject matter	30
b. Terminal understanding of concepts	26
c. En route understanding of concepts	29
d. Certainty of learning	20
e. Amount of learning	21

Abedor developed a quick-scoring technique which enabled a professor to isolate the major problems in the SLATE as soon as the criterion tests and rating forms were handed in. He placed a transparent overlay which showed the desired direction of student response and the cutoff point over each rating form. If he saw a 3, 4, or 5 when a 1 or 2 was desirable, he would add to a tally next to the item number on the plastic overlay. The criterion test would be scored in a similar fashion and when Abedor finished, he knew from criterion test results how many students did not give the desired answers on certain questions, and a quick scan would show what seemed to be wrong. If for example, many students could not identify a certain breed of cattle and also reacted to items 11, 13 and 15 in a way that was cause for a tally mark, one might guess clarity was the problem.

Summary

A rating form can pinpoint possible sources of difficulty. When combined with results from a criterion measure, an evaluator may have enough evidence to begin to hypothesize about why a program results in the achievement of some objectives and why it fails to help students achieve other objectives.

* * * * *

Rating Forms, in Brief

A rating form...

...can yield many useful ideas from a large number of people.

...should be integrated into the usual course.

...should contain specific content and criteria.

...should show what changes are to be made.

CHAPTER XI

Tool Number Five: The Interview

A discussion can yield many more precise ideas about people's opinions on a subject than simply asking them to respond to a question. In a discussion, one individual or groups of individuals can be interviewed by formal or informal methods and an evaluator can probe each person's answers, find reasons, strengths and weaknesses of a project, and seek clarification. It is, thus, eminently suitable as a technique of constructive evaluation. The interview provides the needed link between results and instructional methods to explain why the instruction acted as it did and what to do to improve it.

> Structured interviews can be used with all age groups.

Three-year-old children who view "Sesame Street" have been asked questions to find out what they have understood from parts of a show. A child was asked: "What was the machine's name?" "What was he doing?" "Why did he do that?" "What did he do next?" "Did you like what he did?" "Why?"

Older children who view "The Electric Company" were interviewed about the format used on that show, and asked such questions as "Who is this character?" "What does he do?" "What happens in this picture?" Favorite characters and formats can be identified, with reasons for the choices. In somewhat the same way, film researchers have been interviewing small groups of adult preview audiences to find out if the film is liked, persuasive, and entertaining.

A constructive evaluation interview must be systematic and well planned to provide useful information.

CASE

Applying a Systematic Interview

Suppose there was an instructional unit which needed testing. The unit could be composed of a written portion to provide the basis for knowledge, and a slide and tape presentation of a model of the performance taught, with a practice for the student.

Suppose the objective is to teach students the theory and practice of making a simple animation -- a cartoon. Now suppose that Allan Abedor were to use his rating from approach (as described in the last section) and an interview technique to test the unit:

Abedor begins by selecting a small group of students to help test an instructional unit. He expects some problems in timing and scheduling, and in getting students when he needs them. When he is able, he chooses six to ten students. When the group meets, Abedor tells the group members that the task is to provide information which will help identify and revise the instructional unit on animation. He hands out an agenda and says that the materials, not the students, are on trial. He explains that there will be no revenge for frank, negative remarks, and that he is not there to seek praise or stop criticism.

Next, Abedor tells the students what will happen and what the ground rules are: (see the table of events and rules from Abedor's approach) (1)

TABLE: Events and rules from Abedor's approach

1. Express appreciation for Ss' [subjects'] participation and orient Ss as to the purpose of the session.

2. Relieve Ss' anxiety and facilitate their open and frank interaction.

3. Describe the planned sequence of events, which include:

   a. Pre-test
   b. Individual use of treatment (audio-visual) materials
   c. Post-test
   d. Attitudinal survey
   e. 15-minute "break" including refreshments
   f. Reconvene for debriefing and feedback session

4. Establish the "ground rules" for the session which are:

   a. No talking to each other during lesson
   b. Take notes on type and locating of problems; e.g., don't understand, bored, lesson too fast, etc.
   c. Raise hand for tutorial assistance
   d. Score own pre- and post-tests
   e. Do not cheat
   f. Do not discuss SLATE during the break
   g. Please remain for the debriefing

As soon as all the preliminary student questions have been answered, Abedor begins to follow the planned events. He gives a pre-test, administers the instructional unit on animation, (the text and slide tape), and gives the post-test and questionnaire. During a break, Abedor scores and summarizes the tests and questionnaires to prepare for the group interview.

The group interview is conducted in a systematic fashion to review the work just completed. Its purpose is to uncover problems, find their sources, and decide on possible solutions.

When questionnaires are used as the basis for a group discussion, the diversity of independent student judgments is maintained, and group judgment in the discussion can be compared later with the immediate judgment of individuals. The group interview agenda or, as Abedor calls it, the debriefing agenda, contains test items missed by a certain proportion of students and rating form questions answered unfavorably by a certain proportion of students. If more than 30% of the students, for example, show by their performance on the criterion test that they have a problem in estimating the number of

16 mm. frames for a slow animated sequence, the reasons for failing that test item should be discussed. If the rating form results show that six out of ten students feel there should be more practice on estimating the number of frames for a given segment, then the addition of practice should be discussed.

In addition, if Abedor notices that during the instruction more than a certain number of students ask a similar question, it is discussed in debriefing. For example, if five students ask a specific question during the SLATE, -- for example, how to judge the size of movements from frame to frame, -- that question should be discussed.

Abedor keeps instructional materials readily available for easy reference. When a student says, "I had a problem during the part when the narrator said...", Abedor is able to turn to the spot to locate the exact source of confusion.

The only other equipment Abedor thinks necessary for a debriefing is a blackboard. He lists the problems found in the test, the questionnaire, and the observations. Then the group tackles each problem in turn and develops solutions according to some priorities.

Abedor asks individuals to explain exactly why they answered the test item the way they did, why they answered the rating form in a particular way, and why they behaved as they did during instruction. If, for example, the criterion test shows that students do not know how to gauge the number of frames to depict a slow, moderate, or fast action, Abedor asks the students why they missed the question. He probes to see if the question was poorly phrased or if the students did not understand the principle. He asks students to explain what they did and did not understand about the idea, or asks where they were confused. He might direct the students to return to the spot in the written materials and the slide tape presentation which deals with the principle or provided practice. He might find that the principle was not fully explained and only one example was given; if this should be the case, Abedor and the students might list several solutions before going on to the next problem. The students, in turn, could ask their professor to define the principle on the spot, and perhaps the students could supply additional examples he might use.

If the students' answers to the rating form show that they feel there was not enough practice, Abedor might begin another probe. He could ask, "Where was there not enough practice?" "How much additional practice would you need?" "Would you like the same kind of practice?" "Did the lack of practice make you feel unsure or did it really affect your learning?" He might find that the amount of practice was sufficient but that the type of practice was unlike the behavior required on the test: Before another problem would be discussed, one sample practice would be written out.

To convince the producer of the SLATE that there are problems that need to be remedied, he should conduct or be present at the debriefing. If the producer feels that he cannot carry out the agenda well, Abedor will conduct the debriefing for him.

Abedor tries to take into account and minimize the many factors that can reduce the productivity of the debriefing.

--The interview atmosphere must be open, positive, factual, non-threatening.

--Students should be encouraged to participate and the discussion should be organized around objective data.

--The producer should be taught how to act, and should avoid statements such as these: "I can't be bothered with that problem; you will understand that later." "You read the objectives and you still don't know what they are." Or "You still can't understand the major ideas." Along with an instructor's shrugs and squirming, these comments communicate clearly that he does not want negative comments and blames the errors on the students.

--A time limit should be set for each problem and for the total debriefing.

What are the likely results of a debriefing?

Whole courses may be changed: a sequence of units may be rearranged based on debriefing suggestions. Later units may appear to be better than those created first. Higher post-test scores, less intense debriefings, and fewer problems may indicate better development, better design, increased student ability to cope with the units, or unfortunately, even the students' awareness of the futility of saying anything in a debriefing.

Abedor finds that students are likely to be grateful for being able to have a say in the unit, no matter how poor the instruction, and debriefing is likely to produce more than enough data for a revision.

Abedor expects students to be honest. They may admit how they memorized pre-test answers and breezed through the post-test because the same test form was used for pre- and post-measures. Students may make comments which merely confirm responses made on a test; they will probably give suggestions which may be inappropriate, and talkative group members can monopolize a group discussion. The characteristics of the specific students in a group will lead you to doubt the generality of the information. ("They are all volunteers; the rest of the students won't react in the same way.") Certainly, students' comments are likely to build in momentum and become so overwhelming that the producer gives u

> Abedor says that a debriefing is likely to produce frank comments and defensive reactions. When students believe that their comments will not affect their grades, they can become brutally frank. Abedor expects a producer to become equally defensive. He expects at first that students will test the debriefing leader to see if he really wants criticism.
>
> A producer is likely to become terribly depressed as a result of a frank debriefing. He may wish to abandon the project, or believe it has to be completely redone, or delay his revisions indefinitely. As problems become apparent, the thought of arduous work in the producer's mind is likely to increase.
>
> Abedor believes that instructors may learn to proceed on their own and not make the same mistake twice. They may revise the larger course and get to know the students better. One teacher, for example, who learned more about his students discovered that some of his course goals had nothing to do with the students' professional and intellectual needs; the material was taught simply to please and impress his colleagues.

This interview procedure of debriefing is not perfect: there are some distinct problems. If a debriefing is conducted during a program, it may stop those students who were moving along. The producer may not be able to take notes if he is operating equipment; lights may be interfering or distracting.

Then why debrief if all these problems are present? Because thirty heads are better than one. Students can suggest organization, can sequence, can eliminate extraneous information, can change tests, and can suggest analogies ("a penny is to $10 as 1/1000 of an inch is to an inch"). In Abedor's field experiment, with the use of his model, he secured significantly better results with revised versions of SLATES than he found with original versions of the instructional sequences. Students found the faults and suggested solutions, and the solutions were useful as revisions.

Although project problems can be identified by test scores, attitude survey, and observation, the group interview serves to explain the faults so that sometimes a solution is suggested. And group interviews are relatively easy, inexpensive, and informative.

* * * * *

The Interview, in Brief

Use structured interviews with all age groups.

Plan a systematic interview procedure.

CHAPTER XII

The Test of a Test: Standards for Judging a Constructive Evaluation Test

You can evaluate tests used to assess an instructional project by observing the quality of the results they provide and by gauging their efficiency in providing results. Good test results should reveal the sources of methodological strength and weakness so as to allow for improvement. Good results should be available before it is too late for revisions, and should be collected within the project's resources.

A good test tells a producer what to revise and how to do it.

A good test should be diagnostic.

To show a producer what to change, a criterion test should consist of items that require performance of subordinate skills and knowledge. From the test results an evaluator should be able to see what specific knowledge and skill students have not learned, as well as what they should have learned. The faults can be traced back to the portion of the instruction that attempts to teach those small bits of information, and attention can be given to changes that will upgrade each skill or idea so each portion of instruction can contribute to the total performance.

In addition you can use the results of the test to find out which information or skill is really necessary for the total student performance by correlating the subtests with the final total performance.

Then you can add skills where they are found lacking and, at the same time, reduce the size of your program by eliminating portions which do not contribute to the final performance.

You can build a diagnostic test of this sort from a precise description and analysis of your course goals. You convert into a question each step and decision, each concept and principle which contributes to the final student performance. Each item is constructed so that it can be scored on a pass or fail basis.

An example of the use of a diagnostic test is Gropper's division of a test into multiple choice (recognition) items and construction items. Revisions of his course were made only when students could recognize a [illegible] but not apply it. As a result of one revision the lesson was lengthened from 28 to 55 minutes; performance increased 30%, up to a level of 50%. (1)

> For a producer to be sure he knows what to change in a program when a strength or a weakness is indicated, a test should contain pure items.

Each item should be pure; each item should measure one defined result and allow little influence from extraneous variables. For example, memory should not interfere with concept identification if concept identification is defined to exclude memory. If a child is supposed to identify a concept by pointing to an example among other examples when asked, he should not be asked to recall an example and point to it. Similarly, the test should exclude jargon or notation peculiar to an individual program.

Any student who has mastered the objectives should be able to pass the test regardless of where he was trained. To find out if a test is generally valid you can administer the test to people trained by different programs. (2)

You might also consider the manner in which a question is asked to ensure that all students who know the answer have a chance to answer: reading or listening problems may be interfering with some students' responses.

For a producer to learn all the strengths and faults of his project possible, tests should be broad enough in scope to yield incidental outcomes or unexpected outcomes.

A failing of the narrow test is that it may reveal that goals were achieved, but not that unwanted behaviors may have also been learned. To be as comprehensive as you can in discovering the effects of your program, you must include test items and observer's instructions which will produce reports of effects other than those noted in your goals.

A good measure yields both positive and negative information to tell a producer what to keep and what to change.

Both negative and positive information will increase the likelihood of improvement. If you ask for negative information from students and observers you will get it, although sometimes it can be upsetting.

Negative feedback tells you what to revise; to find it you need a plan and a high degree of self-confidence: everything that one produces has flaws, yet no one likes to be wrong. (3) (4)

Sometimes negative information will reveal extraneous material; students will report what was trivial and what did not contribute to their learning. Other times you will have to extrapolate from students' reports what did contribute to their learning.

Positive information tells you what worked well and provides clues to successful design ideas. (5) Positive information lets you know when you are finished, what to enhance and encourage, what to leave alone, and if your methods are acceptable.

When negative feedback stops, and changes continue to occur which will affect your instructional system, your course has a good chance of collapsing because it lacks the information which tells it to adjust and improve. (6)

Therefore, you must look for information which leads to improvement.

> A good measure is constructed to give insights to the producer as to why the program works and what changes will make the program work better.

The measure should help discover why a particular result appears. Classroom teachers who need this type of insight often ask students to state the reasons for their answers on multiple choice or on rating forms.

Questions can be constructed to provide constructive insights

To inform a producer about what revisions to make, you must phrase test questions so that students' responses indicate a preferred change. For example, a student rating form should contain statements like "More examples should be given" in addition to, or instead of, ones like "The program was boring." One can ask a student to respond to such statements as "The program worked well because________________________," and "Describe the best part of the program and tell why you thought it was the best." "If you could change (or keep) one part of this program it would be__________because______________." But interviewing is a technique best suited for gaining insightful information because an evaluator can probe answers.

> A good test will provide evidence which will convince a producer to make changes.

A producer will be convinced to make changes if a test shows that many students have either achieved or not achieved.

> A producer is likely to be convinced that the evidence collected from a specific test is valid if the test fits the performance requirements of the objective.

To check this criterion you can classify items to see if they fit objectives. And to convince a producer of the validity of the test, show that the test contains situations representative of all the types of situations in which a student will have to behave. The more situations, the better. If two forms of the same measure yield similar results, the measures are probably representative.

A convincing test should have content validity.

The test content must relate to the content of the instructional unit. (7) But remember that some tests don't consist of content at all -- attention measures, for example.

A producer will be convinced if there is high agreement among those who score the test.

If more than one person scores the exam, their totals should be the same. Precise definitions of student behavior (specific objectives) are necessary for agreement. (8)

To be acceptable to a producer, you must show him that the test is not counterproductive.

The process of testing does not counteract the positive effect gained by the instructional method. For example, a test of attitudes toward math should not be so time-consuming and tedious that it be associated with math and influence the students' views.

To convince a producer that the test results are valid, the format and vocabulary of a test should be appropriate to the age level involved.

Students should be able to understand the test question and the possible range of responses. The test should be fitted as closely as possible into a student's normal behavior under the circumstances and a student should have the prerequisites to read and respond to the question.

To be convincing, the test should have face validity.

In some cases, as when a criterion test is needed and students are aware that they are being tested, the test should appear clearly as a test of the subject that was studied. A math test should be perceived as a math test and should not be perceived as a test of both math and reading ability.

To convince a producer that your test results are valid, you need not adhere to traditional test construction rules. (9)

You need not eliminate test items which all students pass or fail: to do so would be to cut off information showing where instruction is good and poor. Standard scores and percentile rank tell where a student stands in relation to a group average, but do not tell you if the students attained the objectives. Keep items that do reflect objectives; eliminate those which do not.

In the traditional sense of the term, reliability shows that the res[illegible]ing scores accurately reflect the ability to perform the task;

thus, a larger test reflects more accurately by avoiding the accidental right or wrong answer. (10) You could compute reliability by correlating one half of the test with the other half, or by testing and retesting subjects on test halves, or on two forms of the test.

If you wish to convince a producer that the test results are valid, then show that the known biases of the test are reduced.

For example, a student's awareness of being observed may cause him to react in the way he believes the evaluator wants him to act. His score may be biased. Unobtrusive measures, random assignment of observation test situations, and placebo observations (beginning the observation with a camera which has no film in it until the students learn to ignore its presence) may reduce the effect of the bias.

To be most convincing, use unobtrusive measures.

Use tests which do not cue the student that his behavior is being observed. The popularity of an exhibition, for example, may be inferred by erosion of the floor tiles in the exhibit area. The number of empty liquor bottles in a trash can is an indicator of a certain level of alcoholic consumption. The degree of fear induced by a ghost story is indicated by the number of children leaving the room in which the story is being told. The size and number of clusters of blacks and whites in a lecture hall is an indicator of racial attitudes. (11) To record unobtrusive observations, an anthropologist constructed a camera which would take a picture of people and objects ninety degrees away from where the camera was pointed.

The use of several measures rather than just one is more likely to provide a sensitive estimate of the effectiveness of a system.

With more than one measure, more errors are likely to be detected, and more of the positive points and the faults of the program are likely to be revealed. Because every project has many facets, using several tests to measure the results of a program is recommended for convincing a producer. The more you test and test well, the more likely you are to be able to understand what happened in a program and explain its results more completely. Because of testing errors and because tests reveal only signs or symptoms rather than actual results, you have to test in many ways to reduce the error.

A problem in measuring many variables is that one measure may interfere with the others. (12) For example: if you stop a student after a segment on which you have measured attention to assess comprehension, you may unwittingly be heightening attention on the next segment. But an evaluator can arrange several measures so they do not interfere. To correct the interference of the comprehension measure, you might introduce filler segments to return attention to normal levels, or test for comprehension on a random basis.

Another problem in using many measures is that students can be forced to spend many hours in testing. To counter this, you can sometimes rank order the tests from easy to difficult, so that when a student reaches his level of ability, you can stop the testing procedure.

Biases should be taken into account.

If, for example, you know that a distractor measure taken on two or more children [illegible] once produces lower overall attention scores than when taken on one child, you can consider an above-average score taken on a group as a good score. If a measure is used in an artificial setting so that you can report most accurately, you can make a comparison between information secured under real and under artificial conditions to check the extent of the bias. If you observe some stable differences between test results collected under artificial and real conditions then you can add some specific quantity to test results secured in artificial situations to estimate results secured in real situations.

To convince a producer that your tests yield valid results, show him evidence that the test has been used and has demonstrated its worth.

Many strategies are available for perfecting tests by tryout.

You could combine an initial tryout of the instructional method or an existing alternative method with a tryout of the test. At that time you could watch students take tests and observe the students' behavior, which may reveal confusing, difficult, and irrelevant parts.

You could confirm the relationship of your objectives to the test by using it on trained and untrained students. You could ask students to complete only some parts of the instructional program and then take the whole criterion test. See if students' test scores cluster according to instructional portions they each completed. You could also test more than one form of a post-test and correlate results to see if they were indeed measuring the same behavior. You could hire reviewers to match test items and objectives to see if they appear valid. (13) Or you could use technical statistical procedures; you could, for example, compute coefficient of reproducibility to verify the test item sequence -- it will predict an individual single response from his total score.

A good test provides results quickly and inexpensively.

A test should be practical -- within the confines of effort, and space resources available. To determine practicality you can ask if it is inexpensive, quickly and easily given and scored, and if the results are useful. (14)

A test should give fast feedback.

It can provide quick information return if it is easily scored and summarized. (15)

A test should be efficient.

A test should cost what you can afford. You should attempt to get the most for your money: you should make tests reproducible. If a test is reproducible, it can provide a common source of results for repeated measures in different environments. Test instructions must be so precise that the same test procedure should be possible under most circumstances. (16)

To make your measures reproducible you must develop the idea, define the properties, clearly state what you are to observe, state rules by which numerals are assigned to the properties of the observed event, and state the condition under which observations should occur.

An efficient measure saves time as well as money; it should be integrated into the program. It should be part of the course procedures, or at least its style is familiar to those who will administer the measure.

You can use professional help in developing tests.

Ultimately, it is most efficient to create your tests correctly; there is then less likelihood of rejecting data because the test was deficient. If you are not an evaluator, you may find merit in seeking the advice of a professional.

A professional evaluator will help (17) plan, develop, try out, and evaluate your measures. In the planning stages he can help you check the logic, the fidelity, the representativeness, and the weights for each objective. (18) Next, an expert will help you develop an item pool, a set of directions, and a scoring system. He will make sure you have as many items in the measure as possible and help you develop more. He will check to be sure that the content of the item, not its form, determines the answer. You and he will start the exam with easy items and end with difficult ones.

He will guide you so that you do not narrow your views too early. Together you can watch students informally, look for trends, then categorize and observe for particular results. He will show you that informality and common sense are more important than rigor in the early stages of constructive evaluation. Later the rigor is necessary when your observations must help you to diagnose and prescribe accurately.

He may know of some standardized tests which you may use to check the effectiveness of your instruction. These tests cost nothing to develop; they have been completed already. Standardized exams are most useful when you are interested in well defined, well understood, tried and true variables but they do not necessarily contain all you are teaching: they leave out important points and contain others you are not teaching at all. (19) (20)

You and a professional evaluator might be able to create a checklist for test selection on the basis of some prime variables such as cost and fidelity. In addition to the criteria listed, the decision to select tests depends on the situation, the attitudes involved, the amount of time a program will be used, the size of audience, the complexity of the program, the cost, and the precedent.

An evaluator will show you how to weigh the criteria used to judge performance. (21) He will also help establish the lower limits of acceptablity for each goal. He will warn you about evaluation pitfalls of which he is aware: a) He will advise you to use small samples for complex measures unless you use item sampling. b) He will recommend that you test for variables in which you are really interested not just for variables you know how to measure. c) He will suggest that you not overemphasize easily defined and measured variables. d) He will tell you to avoid using criteria based on the current conception of schools, which assumes that schools today are satisfactory.

If you are not an evaluator, you can seek the help of someone who is. He can help guide your activities so you will produce acceptable measures. Many educational psychologists are qualified to provide this aid.

> Different types of tests are useful when assessing the quality of drafts at different stages of polish.

Generally rough materials get informal measures, polished materials get formal ones.

TABLE - How measuring tools relate to the degree of methodological polish

Degree of polish of project methods and materials	Measuring Tools
Earliest pre-production drafts --	review by author, producer, expert, concerned person, and other technical staff
Good first draft --	observe, test, and interview individual students*
Good advanced drafts --	observe, interview, test, and administer questionnaires to small and then large groups of students.

*(Used by: 22-27. Led to positive statistically significant results in favor of revised drafts: 28, 29, 30)

Here is a sample combination of measures by stage: In the roughest stage you could conduct a review by author and by an instructional developer. When a good first draft is ready you could administer the rough draft to a few students. When you have a fine advanced draft you could use pre- and post-tests, some informal observation during the course of instruction, questionnaires, and a group debriefing based on post-tests and questionnaire results.

Standards for Judging a Constructive Evaluation Test, in Brief

A good constructive evaluation test:

- tells a producer what to revise and how to do it
  - is diagnostic
  - contains pure items
  - broad enough to measure unexpected outcomes
  - reveals positive and negative information
  - gives insight as to why the program works and how to improve
- provides convincing evidence
  - fits performance requirements
  - has content validity
  - reliably scored
  - not counterproductive
  - is age-appropriate
  - has face validity
  - need not adhere to traditional test construction rules
  - has reduced test biases
  - is unobtrusive
  - uses more than one measure
  - accounts for biases present
  - has been used and found to be worthy
- provides results quickly and inexpensively
  - can give fast feedback
  - is efficient

You can use professionals to help you in developing tests and to help in deciding which specific tests are to be used for different stages of development.

CHAPTER XIII

Supply Number One: A Prototype Unit

It is not economical, nor is it wise, to use constructive evaluation procedures to test a fully produced instructional program. When you have a small portion of your instruction in early form, constructive evaluation procedures are appropriate. If you have a whole instructional program in polished form, you should test it by summative evaluation: only a small proportion of constructive evaluation time is used to test polished final drafts. In other words, the appropriate unit to be selected for a tryout is an early draft or prototype.

A prototype is a model of a larger construction: it has all the parts, but is miniaturized. An instructional prototype is used to teach you and your producers about what affects students. It often consists of a unit: a chapter, a lesson, one of a series of films, or one of a series of T.V. shows. It is to be tested, analyzed, and discarded, as any writer treats an early draft.

A good prototype must resemble the final production.

The attributes of the final draft must be present; it is not necessary for all the rough spots to be polished, but at least they should be there. This may be an argument for not using storyboards (drawings and script depicting an audiovisual) and scripts: they may lack attributes of the final draft. The closer you can get to the final form in an early draft, the better your prediction of the effectiveness of a final draft.

Check all components for minimum technical quality and check to see if instruction is likely to be administered as it is supposed to be. When administered, early forms of an instructional system may not have the smoothness and slickness necessary to stimulate students' interest and attitudes as well as a polished final version. But students can learn and recall what they have learned from early drafts. If you use materials lacking in content (the introduction and sense of continuity are missing) or use a presentation technique which is technically poor with conspicuous defects (smudges of film) (1) (2) you can expect to get similar learning results to that of a final draft, but your motivational results are likely to be off.

The safest prototypes include many formats.

The producer avoids putting all his educational eggs in one methodological basket: if he creates alternative ways of teaching the same things, he should produce a draft containing the use of many teaching approaches. After testing he may only have to eliminate some parts, repair some parts, keep some as they are, add ones like those which are found to be successful, and try new ones.

CASE

Using Many Formats

"Sesame Street" and "The Electric Company" are excellent examples of the magazine format. If, when tested, the data shows one of the segments is so ineffective to be beyond repair, the producers haven't lost everything: they have a dozen others to fall back on. For example, the "Sesame Street" show must have dozens of ways to present the alphabet which employ animation, live action in the studio, Muppets alone, Muppets with children, live film of real objects, fantasy objects, a story line, a lesson, and so on. If they found that when a Muppet and a child recite the alphabet together children attend and practice the alphabet, they would keep the segment and repeat its format. If they found that an adult presenting a "lesson" about letters lost the children's attention, they might explore why; if they found it to be a function of the method, they might abandon that approach. If they thought it had to do with the character, they might experiment for a while with other characters before they rejected the format.

A good prototype is lean.

A good prototype contains only that material which teaches or motivates. A good example of a lean program was Markle's First Aid Course described in the overview. Material was added only when data showed it was necessary.

Fat, the extraneous material which adds nothing to the functioning of the instructional unit, is hard to lose once it's there, but it's not impossible. (3) (4) One method for removing material without increasing the error rate is relatively simple: remove or black out portions thought to be necessary and test the students after administering the instruction. This technique, known as the CLOZE Technique, is also used as a measure of readibility. (5)

The instructional approach in a prototype should be constructed so that a fault can be spotted.

The structure of an instructional system can help provide evidence of the need for constructive evaluation. (6) If the course calls for overt responses at times, the information can be used as evidence. In many systems active practice is required. At those points test-like practices can be inserted, and what might otherwise be invisible mental practice may be observed so that student responses can be analyzed later.

A prototype should be manipulable.

Producers are likely to resist change of a more complete, polished version (7) which has taken a lot of time and money to produce. (8) Therefore, you must ask, "Can the method be easily restructured?" For example, film is less manipulable than videotape, and written material is more manipulable than videotape. Written material on cards is more manipulable than on paper; cards can be reshuffled easily. The greater the manipulability, the more quickly the revisions can be made.

A prototype must be economical.

A prototype is a draft, something to be discarded once it has been tested. No one, except those extremely dedicated to the notion of constructive evaluation, wants to let an expensive draft go.

For purposes of economy many instructional film producers and television commercial producers test their ideas by creating inexpensive versions of their film messages using minimal sets and local talent. The film makers may use 16 mm. film instead of 35 mm., or videotape instead of film: spending extra money for a special nuance of the voice or a particular visual display may not be worthwhile, especially if the experiment may be a total loss. If you are really experimenting, you may be spending a lot of money for no return at all. For this reason it is not advisable to make too many multiple copies of a rough draft.

> If you are going to spend the money to create a prototype, you might as well select a unit which is important.

Select a segment which will provide the most instruction to the most students on some high priority objective. This way you can get the most use from your resources.

> A good prototype should have format, method and other characteristics in common with other units to be created.

The units should be so similar that the results from testing one in the early stages should apply to others. This can save considerable time and money later on.

> Choose a complete prototype.

The unit selected should represent all the methods described in your instructional specifications. The more complete and detailed the unit tested, the fewer the number of tryouts necessary.

You may have some reservations about the validity of test results found on one prototype unit. On the one hand, one can argue that research results on an isolated segment are biased because the results may be different when the segment is embedded in the rest of the program; on the other hand, one can also argue that each unit is likely to be used separately. The essential idea is to pick a large enough unit so that the effect of the unit will predict the effect of the total program.

> A good prototype is hard to find quickly.

CASE

Finding a Prototype Quickly

Allan Abedor and Normal Bell of Michigan State University have developed a method of producing a prototype unit quickly. They set a deadline, and then make the act of planning the prototype a natural endeavor for the producer. They want him to produce a unit of the type he is used to producing, and then help him convert it to another medium if necessary. If he is used to writing, he writes; if he is used to speaking he speaks.

The type of unit they produce has slides and tape, but their procedures apply to any instructional method which has audio and visual components.

Abedor and Bell ask a producer to prepare, by a certain date, a rough outline of a lesson which will meet an instructional objective. They ask him to be prepared to make his presentation to one person. When the producer brings in a lesson on geometry, for example, it is reviewed briefly by Abedor or Bell to find out what will happen during the lesson, and to remedy any obvious defects such as the lack of student practice. When the producer presents his geometry lesson, 35 mm. slides are taken of any drawings, three dimensional models, or other visual aids, that are essential to explaining the geometry principles. The producer's voice is taped; the tape recording is transcribed and edited. Then a professional announcer records the revised script. Students receive a copy of the script, a content outline, a lesson objective, and a number of study questions, and attend to the slide-tape presentation which is the prototype unit. (9)

A prototype unit is not an end in itself, but a means of providing material which can be tested. In the example given, the tapes and the slides made may or may not be used in subsequent units, but they provide material that the producer and an evaluator can examine together in an attempt to improve the lesson.

The results of prototypes will help you estimate the success of a final draft.

Do early drafts really predict the results of a final production copy? Some prototypes do. Many educators, film producers, and television researchers believe they do. Tests of storyboards (drawings representing a film sequence) have successfully predicted audience reaction to films. (10) (11) (12) (13) Scripts work too. For example, in a course, brief written descriptions of problem situations were read to students to see which situations would generate discussion. The problems that did produce discussion were made into films which, in turn, also successfully produced discussion. (14) As to television advertising research, Gerald Lukeman, President of Audience Studies, Inc., said:

> "We have compared the tests of 160 finished commercials against their 'rough' counterparts. The correlation on the average was .90...which means that the 'roughs' are superbly predictive.

Richard Tousey, Vice President of Ramel Film Productions, added,

> "Or, if we may borrow someone else's slogan, that means, that test commercial results are nearly 99 and 44/100% pure." (15)

A good prototype will give you results that resemble a final copy.

Summary

To be ready for a tryout of an instructional project, you must have prepared a prototype unit to test. A good prototype resembles the complete final draft, employs many instructional approaches, includes only what is necessary, calls for continuous overt student response, contains parts which are changeable and inexpensive, deals with important ideas, and includes characteristics common to other units in the project. It is the material which will be examined for its strengths and weaknesses.

* * * * *

Choosing a Prototype, in Brief

A good prototype...

...resembles the final production.

...includes many instructional approaches.

...includes only what is necessary.

...calls for continuous overt student response.

...contains parts which are changeable and inexpensive.

...deals with important ideas.

...includes characteristics common to other units.

CHAPTER XIV

Supply Number Two: A Sample of Students

In most cases, a project director is thinking of a particular group of students when he designs a project. When he puts a prototype instructional unit on trial, he has to get an idea of how those students will react, and therefore, he must select a sample of those students for a test of the project.

> The sample must include individuals who have the characteristics of the target population.

The sample of students should fit the picture of the target population. The students should have the prerequisite abilities, attitudes, and beliefs which define the target group. This implies that you need not choose a sample representing _all_ age groups, or _all_ socioeconomic groups: you should choose _only_ those people representing the prime target group.

CASE

Selecting Individuals From a Target Population

If Abedor, in his work on SLATES, had been testing a remedial unit directed at students scoring below average in agricultural studies, for example, a sample of students scoring average and above average in agriculture would have been superfluous. But he was interested in the general student population of an agricultural college, so he sampled abilities in all three groups. He selected equal proportions of average, below average, and above average-scoring students. (1)

You should consider selecting, from within the target population, a sample of subjects who have characteristics which will help the data-collection process.

Subjects who like to cooperate and are willing to express themselves, for example, are ideal for helping to find the strengths and weaknesses in a program.

The choice of sample should be such that you can find answers to your evaluation questions.

You should feel sure that you can get information from the sample of students which will show the strengths and weaknesses in the program.

CASE

Selecting a Sample to Answer Evaluation Questions

If you want to know if slow readers will benefit from "The Electric Company" television show, you should pick a sample of children who are slow readers, not non-readers. But how many slow-reading children do you need to answer the question. Should you split the sample and give one half of the group one program, and the other half no program? Should you consider other personal characteristics of slow readers which may help find the strengths and weaknesses of the program?

The smaller the size of the sample, the less you can rely on the information.

The size of the sample chosen depends on the degree of generality and inference you want in answer to your evaluation question. The larger the sample, the more varied the information. The larger the group, the more convinced a producer will be of the authenticity of the results because of the possibility for agreement among different students. The greater the number of students questioned, the greater the number of detailed ideas you can get for improvement. (2)

The smaller the sample, the higher the likelihood of getting results which show that the program will not teach the target population when in fact it really can, or will show that the program can teach when it really can't. (3)

Educational researchers often consider 30 subjects an adequate sample. The reason for choosing this number is that the distribution of a group of this size is likely to begin to approximate a normal distribution, and may represent all parts of a given population.

You must take into account your costs in selecting a sample.

There are a few ways of saving time and money in choosing a sample. As you examine your tests by trying them out on a group, you have an opportunity to discover which people may or may not fit precisely to your audience. (4) (5) You can select a small sample and attend to only a few of the most relevant population characteristics. (6) When there are many tests and access to a relatively small number of subjects you can use complex technical procedures and sample among people and test items to draw inferences about whole populations taking all the items. (7)

You should select the smallest sample possible. A tryout with a large sample may provide reliable information, but may cost more. So you may be forced to choose a relatively non-representative sample which is easily available because of the expense of securing a more representative one. You may trade the reliability of generalizations you could make about a population for the possibility of saving enough money to conduct a second tryout.

> Your final choice of a sample is related to the nature of your project and your belief about what constitutes convincing evidence.

If you and your producers believe that information gleaned from an in-depth observation of a few subjects is equivalent to the information received by a superficial test of many, you may choose a very small sample and do extensive observations and in-depth interviews with each subject. If you and your producers believe that the information you need can be asked of a group, that the data collected requires little interaction with students, and that a large number of students is required to find true weaknesses, you may choose a large sample and use criterion tests and questionnaires.

> If you want to find out how you can improve an instructional method, you will want to be certain that the results you get from a particular sample are due primarily to that instructional method.

You may think that you should assign some students to the program and some students to an alternate, but harmless program, or that you should assign some to receive no program. But this sort of experimental design is usually not necessary for the purpose of constructive evaluation. You are trying to collect information which will help you improve a method. You are not trying to convince anyone that this method works better than no program or better than an alternative program: you want to find out which objectives were reached and which were not and you want some hunches as to which parts of the program influenced what results.

To discover the hunches you need to improve, you need only one sample of students who will receive the instruction. You can cross reference different sources and types of data, use logic and theory, and apply common sense to collect enough hunches which will result in a demonstrably better program.

It is usually apparent to anyone that the students' reactions and performance are directly related to the program. Pre- and post-test differences are usually pretty convincing, and attitude questions need no added support to link them to the method. When 20 students, who did not know one cattle breed from another, are able to classify 10 types of cattle by breed, and consistently miss only three after a 30-minute instructional program, most people would be convinced that the instructional program was the principal factor contributing to this change. It is simply not credible to think that over a 30-minute period without an instructional program, 20 students would suddenly acquire knowledge about ten breeds of cattle and be somehow magically misinformed about three other breeds.

Systematically planned tryouts conducted with a sample composed of a few individual students can save much time, money, and effort if used at an early stage of development.

Testing instructional material with a single student can often spotlight a necessary change, one that is easily made in the early developmental stages, but which would be very expensive to modify later. The procedure used to test a program on one student at a time is called the Tutorial Technique. A typical sample might be one student of high ability and one each of average and low abilities. (8)

In the tutorial technique the single student can provide unique kind of information.

During a tutorial tryout you can identify which sections of the instruction are contributing and which are superfluous to a student's performance. You can also coach a student to identify errors within specific sequences of instruction, errors that may not show up in large group tryouts. You can discover, for example, that students are getting the right answers for the wrong reasons.

How can you obtain these types of information? Laboriously. Why? Because as Susan Markle, an instructional researcher, suggested, (9)

> "There are no rules for empirical testing. You are an individual and your student is too, and the situation is essentially a clinical one. If you let your

> first student work by himself while you watch and stay out of his way, you will lose some data. When you question him later, some of the problems will have slipped from his memory. If you talk to the student as he goes through, you need either a fantastic memory or a rapid shorthand for taking down everything that goes on; otherwise you may teach more than you realize and forget later the on-the-spot orally-given frames that produced success. A tape recorder might help." (10)

Fortunately, there are a few techniques and principles to use in conducting a tryout. The following techniques are designed to increase the quantity and quality of the information obtained.

The tryout student should be convinced that he is testing the instructional material and that the material is not testing him. This is a particularly difficult point to get across, since it runs counter to students' educational experiences. As a general rule, the older the student, the more likely he is to react as though he is being tested. This is dangerous because he will tend to criticize himself rather than the instruction. He is also unlikely to volunteer anecdotal information, since this would emphasize and make public what he perceives as his failures. It is usually not enough to tell the student that it is the material which is being tested. He must be reminded as he goes along. This can often be accomplished by such comments as "Remember, we want to find out what is wrong with this material" and "This material needs a great many changes."

If the producer is conducting the tryout, he must also remind himself that the material is on trial. All too often, subtle barbs escape the lips of the author, comments which tell the student that he, the student, is on trial. A seemingly innocent remark -- "I'm surprised that you're having trouble with this question" -- can be interpreted by the student as a statement that he is at fault, and the instructional material is fine just the way it is.

An assessment of the student's abilities should be made before the tryout. One purpose of a pre-test is to determine whether students have the necessary skills to begin the instruction. A pre-test should also focus upon the desired instructional outcomes, the skills which indicate that a student has mastered the curriculum objectives. The arithmetical difference between pre-test and post-test scores may indicate the effectiveness of the instruction. In some cases, where the pre-test would not be included in the finished instructional product, and where taking the pre-test would serve to help or instruct the student, the test may have to be disguised or given at an earlier date.

If the person conducting the tryout has been involved in producing or planning the instruction, this information should be kept from the student. Otherwise, it may prejudice the student's criticism, positively or negatively. The student may also pay more attention to the reactions of the person presenting the instruction, and less to the instructional material.

The student should be encouraged to think out loud, to describe the decisions he is trying to make, to verbalize the mental process. Such information may not only indicate what should be changed, but how it should be modified. To do this, it is sometimes appropriate to interrupt the student. A puzzled look, long pause, question, wrong answer, or a right answer that you suspect might be given for the wrong reason, are all signals indicating a place to stop and find out what is happening. It is often necessary to ask probing questions: "Which part of this problem is giving you trouble?" "What words don't you know?" "What part of the graph doesn't make sense?"

It is important to make a permanent record of all information relating to revisions. If you have to make a change in the program, don't launch into a 20-minute lecture; do record your revisions by writing them on the student's copy.

CASE

Getting Results With the Tutorial Technique

By using individual students to test drafts of a prototype of a programmed text on English money, Rosen, a doctoral student, found that he could, on the basis of test errors and comments of one "bright" sixth grade student, make a revised second draft, and, on the basis of one other student, could make a revised third draft. When Rosen tested the three versions out on three groups of matched students, he found that the two revised versions were significantly better than the original draft, but that the third was not much better than the second. (11)

> The greater the number of tryouts with individual students early in development, the greater the likelihood that the instruction will work and work well.

Individual tryouts cannot go on forever: when two or three successive sessions have shown that target population students can perform according to the objectives without help from the person conducting the tryout, discontinue your tutorial tryouts.

CASE

Ending Tutorial Tryouts

Silberman and Coulsen, (12) educational researchers, used the tutorial technique to test a sample of individual students studying from programmed texts in reading, arithmetic, Spanish, and geometry. The tutor would intervene when a student said he had a problem, or when he looked puzzled, or made an error. The tutors kept records of those problems encountered in the program where assistance worked. Their explanations were worked into the program as revisions. When Silberman and Coulsen felt that a student could proceed unassisted, the original and revised versions were compared. The tutorial testing ended when the revised version was better statistically than the original one, and did not take much additional time.

Here are some examples of the changes in their Spanish program:

> "Items were added to the program in order to provide more practice on difficult structures. A much slower build-up in task complexity was provided, especially in regard to writing in Spanish, in which student performance was consistently lowest."
>
> "Students had great difficulty in following program directions. Steps were taken to reduce the excessive variability from item to item in required response behavior, which was a major source of the difficulty. Other steps taken included simplifying English instructions with symbols, and presenting directions on tape immediately prior to presentation of stimulus material."
>
> "Instead of introducing new Spanish words by dividing them into syllables for initial practice on each syllable, new words were introduced as a unit. One confusing exercise was eliminated."
>
> "Originally, the student would hear a new word once and imitate it immediately, then hear it a second time and repeat it again. This was revised so that a student would listen to new material three times before speaking. Subsequently he would repeat it three times after the Spanish model."

A large group of students can provide convincing evidence for a good draft.

Because producers know that a group test of a project will average out idiosyncratic student responses, producers are likely to be convinced about the validity of group test results. In addition, large group test procedures are familiar to anyone who has attended school.

Procedures for securing data from a large group sample are relatively simple.

If you were gathering data of the interest students pay to an educational film and were going to use criterion tests, rating forms, and observations, you might begin by explaining to the sample of students that they will see a film and be asked questions about it. Then you would show the film. You might keep some lights on so you could observe the students and take notes. You might, for example, count the number of students looking at the screen at given times, or you might ask them to stop the film with a question if they don't understand. You may have to stop the presentation yourself if you see that students who have been asked to respond are reluctant to interrupt the presentation and ask questions. It is not a good idea for students to save up questions until a presentation is finished. If students save their questions they probably do not learn as much, but they also do not help you pinpoint program faults. (13) When the film is finished, you would hand out the test and the rating form and ask students to answer the questions and hand them in.

The choice between a sample of individual students and a sample of large groups depends on generalization, relevance, and practicality.

Consider the compromises in the table below:

TABLE: Relation of individual and group tryout procedures to factors used to choose a sample.

Questions	Use of individual students as samples (Tutorial Technique)	Use of groups of students as samples
Will the sample fit the target population for purposes of generalization?	-The number of test subjects is so small that the results are easily biased. -The number is too small to fit a normal distribution.	-The large sample size helps reduce bias, but it pays to verify student characteristics in the sample.
Will the information be relevant to the central question?	-A tutorial can secure candid reactions and in-depth information. -The tryout style is unusual. -The instruction is not like the real use of the program. -You can distinguish between program and student errors. -You are likely to find motivation problems and not learning problems: students will say what is interesting, but not what is educational. -You will not find oversimplified and inefficient instruction.	-A group tryout is a real use of the program. -It provides greater possibility of confirmation among students. -There is less possibility of in-depth data unless a subsample is interviewed and extensive measures are used. -Subtle errors are likely to elude you. -Bias may become contagious in a group. -Problems may be identified, not solved.
Is the tryout practical?	-It is costly and time-consuming, and requires an expert. -Revision and retest can be done on the spot.	-It is relatively economical for the amount of information secured. -It is sometimes as easy to get a class as it is to get an individual.

You may select students or intermediaries.

You select a sample of students by finding the characteristics which make up the target audience, and then by finding a group of people with the same characteristics; but, on occasion, you may choose a group which is not representative of the target population for a tryout; these people may be called intermediaries if they have something to do with delivering the methods and materials to the students: these may be teachers, parents, administrators, or curriculum experts. Often it is crucial that intermediaries know how to administer the instructional method and materials and be in favor of using the approach. In such cases you must choose a sample of intermediaries.

CASE

Choosing Students and Intermediaries

The Far West Laboratory for Educational Research and Development developed an educational idea called The Parent/Child Toy Lending Library. They produced a series of toys which, when properly administered can be used to stimulate the intellectual abilities of children between the ages of three and four. The program includes a course for parents, a toy library, and a course for teacher-librarians.

There are eight toys (sound cans, color lotto, a feely bag, stacking squares, wooden table blocks, a number puzzle, color blocks, and a flannel board) and forty learning episodes to accompany the set. There is a handbook for parents, a librarian's manual, eight filmstrips and tapes which demonstrate 20 of the learning episodes.

Parents have a chance to observe a demonstration of a learning episode, practice a behavior which may encourage intellectual growth (using exact and precise language, using positive comments, using the child's name, approaching discipline as a learning process and using discipline in a positive way), role play a learning episode with other adults, discuss some educational topic with other parents, and take home a game and use it with their children. After the course is completed, a parent can check out toys from the library.

The product's primary objective is to promote intellectual development. To accomplish this objective parents have to become more competent in helping their child learn, learn to feel that they have a say in the education of their child, and begin to understand what their child can learn. As the result of parental participation the child should become more competent. To aid in the process, the toys have to appear as valuable educational material to the parents, must maintain the parent's interest, and be easy to distribute and handle.

Tests for each program element -- the toys' features, parent behaviors, and child behaviors -- were created. For example, experts reviewed the toys using certain criteria created by research staff at the Far West Lab. An observer watched to see if a child wanted to play with a toy after five sessions of 10 - 20 minutes each to gauge his interest. A satisfactory toy was one that maintained interest for 80% of the children after five sessions.

Parents were asked four open-ended questions:

1. What did you learn from this experience that was useful?
2. What was the most interesting part of the experience?
3. What didn't you like about the experience?
4. How would you improve the program?

Children were tested on the Responsive Test, a test used to measure intellectual achievement. Far West staff chose samples from three different audiences: educators, parents, and children. An available sample of people to be reviewers of the toys were chosen from staff researchers. The parent courses and the toys were tested on parents of particular children in four places: Berkeley, and East Palo Alto, California, and Murray and Jordan school districts in Utah. The sample of parents provided people with characteristics of those considered to be the population of program intermediaries. Parents from East Palo Alto were primarily black working-class; from Berkeley, white middle class; from Jordan and Murray, white and Mexican-American working-class.

Summary

The sample selected for a tryout must reflect the target population, must help in answering the evaluation questions, and must be practical. Large and small group tryouts are useful for different purposes. A small sample tryout is most appropriate in the earliest phases of development, while a large sample tryout makes most sense when a good prototype is ready.

The next element to choose is a setting for the tryout. After tests, prototypes, and samples have been chosen, you pick a test site.

* * * * *

Choosing a Sample of Students, in Brief

Choose a sample which...

...uses students or intermediaries.

...will help you answer evaluation questions

--from target population.

...will be practical

--within your budget.

...will provide convincing evidence

--large enough to rely on (generalization)

--but appropriate to the product's phase of development

small samples early
large samples for a polished product.

Convincing evidence procedures are simple.
A single student can provide unique information.

--certain results of test are due to the method.

CHAPTER XV

Supply Number Three: A Test Site

As an instructional project is taking shape, a project director must take into account the place in which the teaching method is to be used. It could be at home, in an elementary school classroom, in a large auditorium, in a room with twenty movable chairs, in a laboratory full of equipment, or in a library with carrel facilities.

> To make the best use of your resources, and to increase the likelihood of completing a tryout successfully, the choice of a site must be practical.

You must be sure you have a site located near people selected for the sample, enough staff to cover the number of test sites, enough money to cover cost of equipment to be used and transportation, and a place large enough for the number of people, the size of the equipment, and the nature of the program.

> When you prepare for a test of an instructional project, you must choose a place for the tryout much like the one in which the method is most likely to be used.

The test site should simulate the instructional setting where the method will be used; the closer the representation, the more generalizable the results to classrooms with the same attributes. But you may choose to represent only some of the characteristics by using an artificial setting: a plain room, for example, with chairs and a blackboard instead of a real classroom. You may even represent the real setting on all dimensions by a field test in one of the places in which the instruction will be used -- a real fourth grade classroom, for example.

> The events and objects in a test site must be controlled so that you may feel a degree of confidence that factors other than the instruction did not make the change.

Because you want to know if students from a certain group learn from a certain method in a certain setting, you might control setting variables to be sure that no unrepresentative feature of the setting has a significant effect on the instruction. You may have to caution a teacher about changing the physical setting in ways which may influence the most important results of the program: posters, books, teachers, class size, or instructions may alter the effects of the program: when the program is tried again, in a setting where a teacher follows your method to the letter, the results may not be duplicated.

> A laboratory test site provides the control necessary to discover precise, but not necessarily generalizable, answers to evaluation questions.

By standardizing a setting, for example, by requiring a test to take place in a certain room with only certain features, you exert control. When you finish a tryout in a controlled setting you can usually say that what resulted was due to a specific method. But controlled conditions are often artificial, and any artificiality prevents you from promising a person in an uncontrolled environment that he will get the same results.

CASE

Using a Laboratory Test Site

The shows "Sesame Street" and "The Electric Company" are often tested in a laboratory test site. C.T.W. researchers take distractor equipment to measure the distraction scores of a show -- television, videotape player, and rear view slide projector -- to a school. One child at a time is observed. At times researchers may stop the tape and ask the child what happened and what will happen. Even though the tryout takes place in a school, these sites are considered laboratory settings because the environment represents some facets of the natural viewing situation (the natural distractions are represented by slides) and includes interference with the instructional method for purposes of testing (the observer's questions).

When Milton Chen did his research on the verbal responses of children to "The Electric Company", he went to viewing centers and schools. He was observing situations in which the show is usually watched. But he did interfere in the natural setting somewhat with the presence of observers, tape recorders, and hanging microphones. Although Chen's evaluation took place in the field, his interference introduced a characteristic which might have been responsible for some results, and reduces his ability to generalize to other such places. Therefore his test site may be called a laboratory. (1)

Although the results of a laboratory test must be qualified, researchers have been successful in predicting field results by using results gathered in an environment which partially represents the real one.

CASE

Predicting Field Test Results

The Communications Research Group at Dupont has used laboratory test sites for improving the teaching ability of their commercials. A typical laboratory test would proceed as follows. To test a commercial for Lucite paint, the researcher selects 60 homeowners who painted some part of their homes within the last two years and who watch at least two hours of television per day. First, the researcher tests the homeowner's attention to the commercial. He shows each subject a 20-minute film in which the test commercial and other commercials are embedded. The viewer controls the degree of screen brightness by pressing a foot pedal. His presses are recorded and subsequently scored. If a subject stops pressing the pedal, the picture becomes very blurred but is not completely gone. Slides of outdoor scenes are projected within view of the subject; these slides act as a distraction. Each subject is told to choose to look at or ignore the television depending on his interests.

To measure learning under optimal motivation, researchers tell the viewer to look at the commercial as many times as he must to learn everything he possibly can; if he can remember a great deal, he will receive a reward. He must still press the foot pedal to see well.

The viewer answers a self-administered questionnaire in which he tries to recall all messages. To arrive at a scoring procedure for learning, a team analyzes the commercial message to determine the number of "message links" -- as many of the possible simple facts which can be extracted from the commercial. Examples of message links are a brand name, a product, an event. The commercial writer differentiates between message links which are of primary importance (a viewer must learn these for the commercial to be successful), those which are of secondary importance (these can be sacrificed to insure learning of primary message links), and ones of tertiary importance (these are not necessary for the viewer to learn).

Dupont researchers are able to use the scores derived from a laboratory test site to predict results gathered in a more natural field test setting. One field test at Dupont consists of telephone interviews in which a subject must prove he saw the program by recalling key program content before and after the commercials. Then the viewer is asked to recall as many simple facts about the commercial as he can. The total score is the number of message links recalled.

To make a prediction of field test results, Dupont researchers combine scores for different measures into a formula. The formula is simple:

Communication effectiveness = (attention level)* x
(recall under optimal motivation)

*Each variable is multiplied by a constant

Attention level is computed by a ratio of foot pedal pressing under unmotivated and motivated conditions. The recall of message links is scored on a scale from +100 to -100. For example, the paint commercial got an attention level score of 80%, which is considerably better than average. The 80% was multiplied by the recall score, 23. Thus, according to the formula, the Dupont researchers would expect a recall score of +18 .40 = (80 x 23) in the more natural situation. That means that when subjects are called at home after viewing the commercial on the air they should only be able to recall the amount of primary and secondary message links which would be scored around +18. The actual learning score obtained in a field test was +20 (out of a range of possible scores from +100 to -100). At Dupont, communication effectiveness of a television commercial is predicted in nine out of ten cases by plugging average scores of viewer attention and recall into the formula. (2)

> A field test site -- a situation in which methods and tests are used precisely as they would be if the instructional method or product were already in use -- provides trustworthy results.

The planning difficulties (travel, teacher education, possible dropouts) of a field test may be worth the inferences you are allowed to make because field test results are derived from a sample of the precise setting in which the program will be used.

CASE

Using a Field Test Site

When the Southwest Regional Laboratory had a good draft of a program ready to teach concepts to pre-school age children, they selected field test sites. Two inner city schools and one rural school took part. S.W.R.L. researchers were willing to put up with the travel, the orientation of teachers, and the possibility of teachers dropping out or distorting the program because they knew the results they could get would be applicable to most of their target settings.

To a certain extent the children's abilities -- the characteristics of the sample -- determined the field test site in this case. The schools were selected because the children's mean scores on a 10-item pretest of concepts fell below 50% correct. Two schools could not participate because of scores better than 50%. (3)

CASE 2

Using Field Sites in Advertising

Advertising researchers use other field test techniques similar to the phone interview; other techniques include cable television, in-home interviews, letters, and trailers distributing redeemable coupons near supermarkets. In this technique a trailer is posted near a supermarket. Customers are invited in and are asked to view commercials. The evaluator gives those who see the commercial redeemable coupons for the product. An equal number of people who have not seen the commercial are given redeemable but identifiable coupons. The evaluator counts the difference in the number of coupons redeemed by those who saw the commercial and those who did not as his effectiveness score. If more coupons are redeemed by those who saw the commercial, the message is probably getting through.

A test made at a field site can be structured so that important features of a course can be taken into account later.

You can make up for the complexity of a course tryout in the field by systematically recording what you observe in different test sites.

CASE

Accounting for Course Features after a Field Test

Richard C. Anderson tested a program in population genetics at field sites. The field test started after all students in a pilot test scored 90% or better on a criterion test consisting mostly of constructed response items, problems to be solved, and concepts and principles to be defined and illustrated. Two high schools participated. The groups consisted of 750 high school students, nine teachers in 30 classes. The teachers were told to use the program according to their own best professional judgment.

Teachers were allowed to use the program as they saw fit, for he suspected that the way a teacher used the program would affect its achievement. Records were kept on use of the program materials and teachers' approaches were categorized in three classes: 1) those who made the program available but did not require completion and did not allow class time, 2) those who required the activity but allowed no class time, and 3) those who gave a definite assignment with up to three hours class time. The percent correct on the achievement test for the first two groups ranged between 45 - 50%; the third group scored better than 60%. Knowing about one source of variability in the achievement score helped Anderson decide why the program succeeded or failed and what to do about it. (5)

There is a good reason for observing and recording the features of a setting and it is exemplified by Anderson's field test. You need a record of interaction of characteristics of a test site with the program so that you can pinpoint the different effects of the setting and the effects of the program. You may find out that a program works in one of your test sites and not in the others: there may have been something in the test site which made the difference. If you can find out what the factor was, the revisions you suggest may relate only to the setting, not the program. For example, one may suggest that the instructions for Anderson's program include specifications for the program to be used in settings where the teacher will require the program and give class time for it.

Test sites may vary from tightly controlled artificial settings to natural settings.

CASE

Sequencing Laboratory and Field Tests

> At the Far West Regional Laboratory the usual procedure is to progress from feasibility studies (laboratory) to studies in the field with no interference. The first tests of the toy library might take place in the offices of F.W.R.L. The second round of tests might take place in a real community, but the Far West researchers would be along. A final field test of the toy library might consist of sending out the materials and test instruments and allowing a toy library to operate by itself.

The table below summarizes the features of laboratory test sites and field test sites and relates these features to the criteria used to choose a site.

TABLE

Criteria	Characteristics of Lab Test Site	Characteristics of Field Test Site
Representation of the characteristics of a real setting	It simulates characteristics which influence learning most.	It may be a good representation if one or more real settings are used.
	It introduces some artificial features to get information.	There are no artificial constraints.
	You have a captive audience specially selected to represent characteristics.	You have a natural audience, which comes with setting. The audience may not fit the characteristics of your target population, but you could pick the setting on the basis of the sample present.
	The teacher is selected, or a real teacher is placed in a mock setting.	The real teacher comes with the setting.
Control of the instructional variables to be able to answer evaluation questions.	You have great control over the program and simulated setting variables.	Your control over the setting is minimal. There can be great control on the program and you can use objective observations of setting variables to take variables into account later.
	You can predict field results.	You can and must account for variables to record more exact program results.
	Some variables are uncontrolled because they are not present.	All variables are assumed to be present.

TABLE (continued)

Practicality	It is costly or inexpensive depending on equipment required.	It is costly and complex, and problems are magnified in the real world.
	It can be used to train your staff.	It often requires a pilot test to train your staff.
	You can use any material for an early check on feasibility of the program.	The high cost of the test at a field site prohibits the use of poor quality materials; you should use a fine draft only.
	It is time-consuming to set up, but easy to administer.	It is time-consuming to set up and administer.

Summary

If you follow the instructions for creating and arranging the elements of a tryout, you will have a complete set ready: measures, prototype units, samples, and test sites. Your measures may consist of a review, a progress test, a criterion test, a rating form, and an interview, and your measures would fulfill certain criteria so that the tryout results would be meaningful. Your prototype units would be fitting for the type of tryout you have in mind. Your sample and setting would reflect your target population and instructional setting, would allow for control, and be practical.

Once you have assembled your tools and samples, you still may not be able to conduct a tryout: the elements must be coordinated so that they mesh and so that the tryout will run smoothly and provide data to answer questions.

* * * * *

Choosing a Test Site, in Brief

Choose a test site which...

...is practical.

...represents the characteristics of the real setting.

...is controlled enough so that the evaluation questions can be answered.

A laboratory test site provides...

...control,

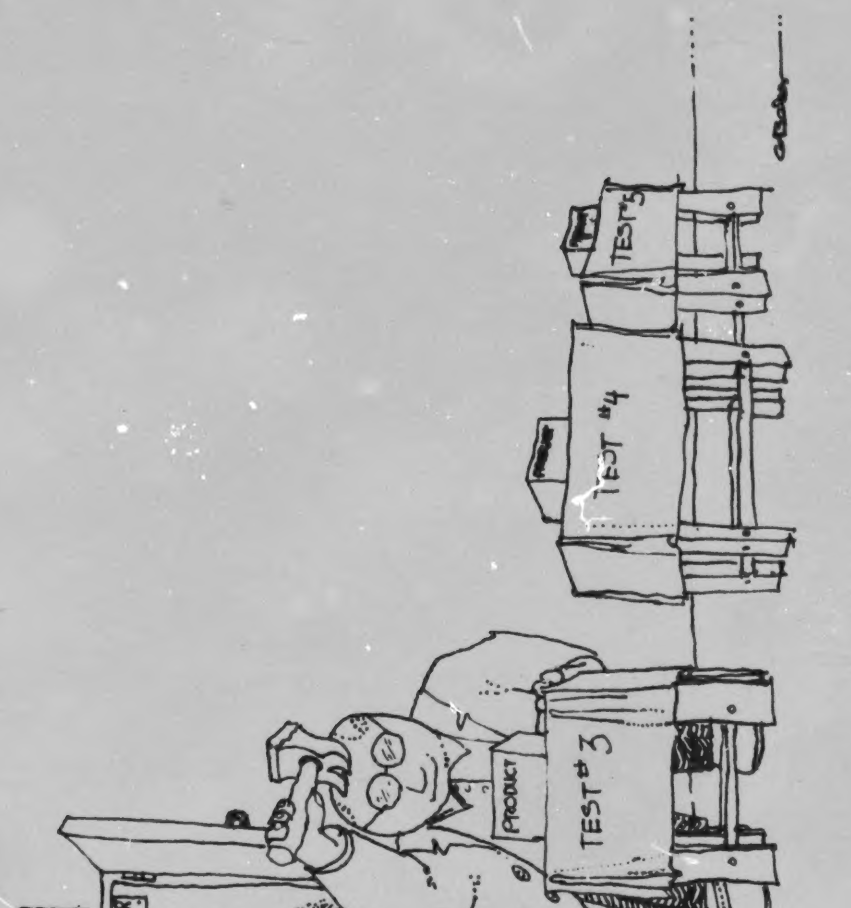
but not necessarily real characteristics.

A field test site provides...

...limited control,

but real characteristics.

...opportunity to take into account course features later.


- OFFICIAL -
TESTS INSTRUMENTS - 1 THRU 5
SLEDGE HAMMER
SLEDGE HAMMER
HAMMER
TACK HAMMER
DUSTER
PRODUCT
TEST #1
TEST #2
TEST #3
TEST #4
TEST #5

CHAPTER XVI

Trial for Error: Organizing and Conducting a Tryout

Mark Twain said "Get your facts first; then you may distort them as much as you please." How do you find the facts? In constructive evaluation you secure facts by a tryout, a procedure in which you secure data to answer your constructive evaluation questions.

> To plan a tryout you decide on a combination of prototype units, tests, samples, and test sites.

By this time in the planning of constructive evaluation you should have decided on the nature of the instructional unit you will administer. You begin making your tryout plans by saying:

I AM GOING TO ADMINISTER THE INSTRUCTIONAL UNIT (AND TESTS) I HAVE SELECTED (choose any number from a. - d.)...	a. as is to a sample of students.
	b. and administer a comparable program to another sample of students (e.g., traditional version).
	c. and administer tests but no program to another sample of students
	d. and administer variations in the same unit to other samples of students (e.g., original and revised versions)
I AM GOING TO ADMINISTER A UNIT WHICH IS (choose one)...	a. a first draft
	b. a rough but revised draft
	c. a polished draft
	d. a final draft

The choice of the elements of a given tryout depends on many things; one of those deciding factors is the quality of the draft you have.

If you have a first draft, you can review and then test individuals in a laboratory test site. Many project directors use this approach to test programmed text materials and the Children's Television Workshop staff uses this combination to test "Sesame Street."

If you have a rough but revised draft, you can review and follow by testing groups in laboratory settings. Group laboratory tests are used by film producers to try out new films in theaters equipped with mechanical responders.

If you have a polished draft, you can review and follow by testing individuals in a field setting. Individual field tests are those used when a few children are asked to respond to a unit in a regular class setting.

If you have a final draft, you can review and follow by testing groups in field sites. The Southwest Regional Laboratory used the group field test for the Concepts Program and so did the Far West Regional Lab when testing the Toy Library.

These are not hard and fast rules; consider them as general guidelines only.

By this time in the planning of constructive evaluation you should have selected a number of tests to use. You continue your tryout planning by saying:

I AM GOING TO USE (choose any number from a. - g.)...

a. review
b. pre-test
c. post-test
d. progress test
e. rating form
f. interview
g. post-test for long term memory or application

Tryouts can be divided into short and long-term tryouts, depending on the nature of the result you desire and the complexity of your program. At present most constructive evaluation tryouts do not run more than a school year. You would continue your planning by saying:

MY TRYOUT WILL TAKE...	a. a short time b. a long time

You continue your planning by saying:

I WILL ADMINISTER THE UNIT AND TESTS TO A CERTAIN NUMBER OF SAMPLES OF (choose a number)...	a. individuals b. small groups (6 - 30) c. large groups
THE PEOPLE (OR CLASSES) IN THE SAMPLE (choose a combination)...	a. are to be randomly assigned to the unit studied they are in (if more than one unit is used) b. are to be randomly chosen from the target population c. matched to other students in other groups based on certain characteristics (prerequisite abilities)
I WILL ADMINISTER MY PROGRAM AND TESTS TO MY SAMPLES IN A CERTAIN NUMBER (choose a number) OF...	a. laboratory test sites b. field test sites

When you have made your choice of a combination of elements, you substitute a specific plan for each general one. You state which units, which particular tests, which population, how much time, and which test sites you intend to use in your tryout. Then you plan the tryout itself.

You should end with a summary statement coordinating the major elements of a tryout:

I have five units (numbers 2 - 6) in a supervisor training program which are rough, but which have been revised and which I will test as is. I will test these by review (R. Scott, A. Porter, and W. Schmidt, experts in the subject) by pre-test (A Test of General Abilities) by a post-test (five simulated problems), by a progress test (adaptation of the distractor measure) and by an interview (conducted by project staff asking for improvements). The tryout will be done over a short time (one month). I will present the unit to ten individuals (five first-line supervisors in T. P. Co. and five of their trainees) and simultaneously present it to one class on first-line management selected at random from T. P. Co.'s five classes. I will ask the ten individuals to think aloud as they go through the program with a production staff member in a laboratory setting (our offices), and will ask the class to participate in a field setting (in their class as the units would naturally be used and without the progress test).

A good tryout plan can be discovered or planned or both.

You can discover a tryout technique by simply observing to see what happens to students. You look without being directed by asking a specific evaluation question; you keep an open mind. But do not spend undue time gathering diverse observations, which often are not put into order and only confirm the obvious.

You can plan by asking specific questions about what you believe should happen. But be careful -- this approach is likely to narrow your view and you may miss some interesting and key discoveries. To be comprehensive you should observe many aspects of behavior to answer a wide range of questions.

Your tryout should be orderly, constrained, and deliberate. (1)

Because you cannot wait for large scale scientific investigations to answer every question (2), you will probably have to use relatively informal tryout procedures. In fact, traditional experimental design used for educational research is generally not useful in constructive evaluation:

> "...the application of experimental design to evaluation problems conflicts with the principle that evaluation should facilitate the continual improvement of the problem." (3)

But informal procedures are not automatically sloppy or nonrigorous.

A tryout can and should be based on valid data.

In instructional evaluation and research it is possible, according to some technical experts, to infer relatively sound causal conclusions without all the requirements of true experimentation.

To do so you must select randomly from an appropriate target population, check the validity of your tests, and control the test situation to the extent necessary to reduce the number of likely explanations for the results. To reduce the number of possible explanations, you can also specify and measure conditions which influence test results such

TABLE:

TABLE 1

PLAN FOR THE IN-CONTEXT TRYOUT OF INSTRUCTIONAL MATERIALS

Prescription Assignment Procedures that:

1. require all students to use the same instructional materials.
2. allow students to select the appropriate prescription.
3. allow students who fail a test to receive a new prescription written by the teacher based on the appropriate cause of failure.

Test-Taking Procedures which:

1. insure an accurate measure of student performance.
2. forbid any assistance from the teacher, aide, or other students.
3. prevent the student from using the instructional materials during the test.
4. require equivalent forms of tests taken after each test failure.

Test Interpretation Procedures which:

1. provide an accurate decision about mastery of each objective.
2. are consistent across students and tests.
3. define tests as the standard of performance.

Classroom Management Procedures which:

1. encourage students to learn from the materials.
2. provide for student decisions.
3. decrease the amount of down time.
4. are consistent.

Student Behaviors which:

1. permit self-scoring of materials.
2. allow students to be self-evaluators.
3. allow students to solve their own problems.

TABLE 1 (continued)

PLAN FOR THE IN-CONTEXT TRYOUT OF INSTRUCTIONAL MATERIALS

Teacher Behaviors which:

1. use reinforcement techniques to motivate students.
2. prohibit student tutoring.
3. provide consistent day-to-day behavior.
4. provide consistent judgments of student performance.

as creating an experimental atmosphere, and telling students they are being tested, by countering them with statements to relax students with unobtrusive, unnoticeable tests and with carefully controlled testing situations.

Actually, many of the variables that evaluators try to control make little difference in the results. Among the few that researchers have found do make a difference is the experimenter's bias, which alters test results. (4) (5) To minimize this kind of bias, test administrators should have no particular expectations concerning the results.

You will know you have good tryout procedures from which to generalize when you collect similar test results in repeated tryouts and when you collect similar test results in different, but realistic settings. (6)

CASE

Controlling the Tryout

In her master's dissertation Judy Light (7) showed that many factors can influence the results of a program in a classroom and thus affect the inferences as to what caused what. However, she felt that these factors could be controlled. She tried to manage teaching procedures, student motivation, and testing procedures. Here is an example of her tryout plans:

These conditions are difficult to maintain and seem most useful for self-instructional materials to be used under carefully controlled situations. Unless the same controlled procedures -- highly rewarding, with no tutoring and tight test procedures -- are to be used in the program under real world conditions, she may find and remedy many learning faults, but in the real system other learning faults may appear. To conduct an orderly tryout, you will need a checklist of things which must be done. Here are a number of items for you to build upon:

1. Acquire permission to use space and students: secure transportation, if necessary, for the participants or for moving equipment or material. Leave enough time to set up the space. (8)
2. Prepare to inform teachers about what you are trying to do and what they are to do. Teachers resist thoughtless try outs. (9) If they say they are concerned about trying out new materials because the materials are untested, remind them that most classroom instruction is subject to the same criticism.
3. Remember to encourage teachers on early trials to develop alternative methods and record them. (11)
4. If the teacher uses the material, tell the teacher to use the material as he normally would.
5. Be sure the people in the sample can get to the test site. Give them a number to call if they cannot come.
6. Include extensive instructions to students. Tell them to state what is confusing, difficult, old, dull. (12)
7. Tell the students to answer all test questions. Tell them that if they change answers, make certain they mark them out, not to erase them. Ask them not to cheat by copying feedback. (13)
8. Prepare instructions which do not bias the student's attention by telling where the material comes from. Include an instruction to students not to pay attention if they find the stuff dull.
9. Prepare to brief students about the ground rules and instructions when meeting; that is, tell the students, for example, that the schedule is as follows: administration of program, first; next, collection of achievement and attitude data; then quickly scoring and tallying results, then collecting the observations of the students during the program; next, developing a debriefing agenda; and, finally, conducting a debriefing. (14)

10. Prepare an informal test to see if students know the ground rules. (15)
11. Prepare to record all questions during instructions.

CASE

Planning Tryout Procedures Precisely

Abedor followed a carefully contrived agenda in his tryouts:

Instructional Development Tryout Session

I. Preflight Facility:

Check software installation and operation in each carrel. Check for required number of workbooks, pre- and post-tests, answer sheets, keys, data matrices, reactionnaires, audio-recording equipment and problem-posting flip chart, and refreshments.

II. Student Arrival:

1. Pass out name tags

2. Create atmosphere of informality and low threat

Students have volunteered for this session and are unsure as to whether this will adversely affect...their grade in the course, future employment, or...other more horrible reprisals. They must be put at ease or very little constructive criticism will be forthcoming. Therefore, wear informal clothes (the student will) and make small talk as students arrive.

III. Introductory Remarks

1. Welcome:

Thank students for their willingness to help you revise your "first draft" materials. Assure them that their frank and honest opinions are of crucial importance and that nothing they say will in any way affect their grade, job, or pose other threats. It is the author and the program which are under the gun--not the students.

2. Role of Students:

To help you identify weaknesses in the materials, procedures, or exams, and to make comments and/or suggestions for improvement. You are looking for comments pro and con on "relevance," "redundancy," "boredom," "obscurity," "clarity of visuals," "needless make-work," "poor exam questions," etc.

3. Role of Author:

   Your role is to gather data and suggestions for revising the materials and to provide tutorial assistance to the students on any aspect of the lesson.

4. Overview of the Procedure:

   The tryout will begin with a pre-test (to assess how much they know to start with); then the lesson materials; then a post-test (to determine how much they have learned from the materials); followed by an opinionnaire and then a break, with refreshments. After the break will be a group debriefing.

IV. General Instructions

1. Test Scoring: Both pre-test and post-tests are self-scoring; students score their own. Please mark incorrect answers on the answer key--not in the test booklet.

   Scores do not count towards a grade; they are for your information and to show us weaknesses in the lesson.

2. Be Honest: Don't look at the answer key before or during the exams. If you artificially inflate your score, we don't really know how good (or bad) the lesson is.

3. Guessing: Guess at the answers you don't know, and place a question mark after your answer on the test booklet. If you don't understand the question, place a question mark in front of the question in the test booklet and the answer key.

4. Ask for Help: If you have problems during the lesson, raise your hand and I will come over. Do not talk to your neighbor.

5. Write Down Your Problems: When you have a problem, write it down in the workbook.

6. Reactionnaire: We need your opinion on several critical aspects of the lesson design. Be frank and honest as you fill this out.

7. Break: Have a Coke and don't go away. We need you for the debriefing.

8. Debriefing: We will reconvene to discuss the lesson, using exam scores, reactionnaire data, and your notes and comments to organize the discussion. Remember, any comments you make will be useful.

> Tryout procedures should be easy and simple to remember and carry out, manageable, and self-explanatory. (16) (17)

The tryout techniques you use should be spelled out so clearly that different staff members at different test sites could carry out similar procedures. If procedures are replicable, the results may be comparable. (18) (19)

> To find out what's going on in a complex course, you will have to choose between the number of activities in a tryout and the risk of failure to secure the information.

For example, if you were using pre- and post-tests, presentation of the unit, observations, attitudes, assessment, and debriefing, and you assigned different people to produce the tryout plans for each part and carry them out, you might get a lot done quickly, but you also risk a certain amount of failure if any one part fails to function well. This is most critical when one part depends on another. To prevent a loss of information because of complexity, a tryout procedure should be refined by a trial run, or by having a team try to find holes in the approach.

To proceed smoothly, staff training is necessary for most constructive evaluation tryouts.

Training should be to produce a staff consisting of two groups: those who can skillfully carry out replicable procedures, and those who can create tests and tryout methods. The staff should be trained to follow rules and to persistently question authority tactfully; questioning authority is not likely to gain friends, but, if done tactfully, is likely to gain respect.

Remember that materials and procedures, and trainers are needed for training. These require time and money.

Study the production phases of your own instructional system and adjust your tryout times to the idiosyncrasies of the production system.

Generally, benefits are greatest when constructive evaluation is used...

> "At the earliest stage in planning at which useful information may be obtained and the latest point at which changes...are practical." (20)

This is not to say that data from final versions could not contribute to the production of later segments.

CASE

Adjusting Tryout Times to Production

In a large-scale repetitive instructional television sequence like "Sesame Street", most of the research benefits can be made most easily during early production times; that is, at script and planning stages. Yet some data is needed on final tapes because material does not jell until it is put into final form.

Schedule a tryout so that it comes either at the time in a course when the material would be taught, or when students have only the prerequisites required. (21) And leave adequate time to make revisions between tryouts.

CASE

Testing at Early and Later Stages of Development

After the five test shows of "Sesame Street" were produced in July of 1969, a considerable amount of field testing for constructive purposes was begun. The main purposes at that time were to test the attention-holding ability and the teaching ability of the shows. In addition, measures were tested and tryout procedures were observed. Children of different socio-economic classes in two cities were observed and tested at home and in day-care centers. This was the heavy testing done at the earliest stages of production. A great deal was learned from this work. (see table) (22)

TABLE: Abstract of the major findings of the five test shows of "Sesame Street" from a C.T.W. report.

"The major findings of the studies reported here may be summarized as follows:

1. Four-year-old children who viewed the five hour-long test shows made positive gains on tests over various CTW goals. These gains appear to be positively related to (a) the amount of emphasis on the specific goal in the programming, (b) the manner in which the goal-related subject matter was presented, and (c) the extent to which the children exhibited relevant overt responses to the given program segment.

2. Background characteristics of the children are related to the average level at which they are already functioning in virtually all goal areas. On pre-tests, children from middle-class neighborhoods performed at a higher average level than children in day-care centers, and the latter, in turn, out-performed disadvantaged children who had had no previous classroom experience. Positive gains were found in all three groups.

3. The visual attention of the four-year-olds was as high for the test shows as for any other children's programs previously tested, including both commercial and non-commercial cartoon and live-action. The research demonstrated the feasibility of sustaining the visual attention of four-year-old children over an hour-long show.

4. Repeated exposures, varied treatment, and visual simplicity (freedom from irrelevant elements) were generally the most effective treatments from the standpoint of instructional effectiveness. Careful manipulation of such factors can lead to significantly increased instructional effectiveness.

5. The tests designed by Educational Testing Service and administered as part of the study reported here have been found by ETS to be acceptable in terms of important technical characteristics, and have been revised as a result of this study.

6. A great deal of monitoring will be required in order to sustain the experimental conditions of "viewing" and "non-viewing" in the case of children studied in their own homes."

Now as the show is being prepared for its fifth year, testing is coordinated with production in a different way. Scripts are reviewed before they go into production, and as production takes place advice is given. Some shows contain new techniques or characters. As soon as a show with a new feature is complete, tryouts are done on small samples in local settings before many more shows are made using the same technique or character. Sometimes the producer eliminates the segment containing a technique found to be faulty; sometimes he leaves it in but does not include it in later shows. The idea is that the information must get back to the source of production -- the writer and producer -- as soon as a prototype is tested. This implies that some planning for a tryout should take place before the prototype is produced so that the tryout can take place as soon as possible.

A tryout should be feasible within your total resources, be relatively inexpensive, be acceptable to classroom teachers, and require few subjects.

You will want to spend more money and effort on tryouts which are designed to answer the more important evaluation questions.

To be sure that teachers will use the program, you must, during a field test, assure those associated with the program -- teachers, principals, students -- that you are not evaluating them. Do not comment on their performance. In addition, do not test students so much that they lose motivation to perform.

Your tryout should fit into your instructional program; it should not be an extraneous piece tacked on.

If possible, tests and observation procedures should be a part of the program as the program will eventually be used. They should, at least, not interfere with the program. In other words, do not collect data in a manner that distracts from the presentation. For example, go easy on record-keeping and taking up student time. When gathering a great deal of information, sample among students and teachers. Let the teachers use the program as they feel they should. Place exercises which can be used as progress or criterion tests into the program.

CASE

Integrating Constructive Evaluation in an Ongoing Program

A good example of the integration of data collection for constructive purposes into normal class procedure is the Individually Prescribed Instruction Project at Pittsburgh. Elementary school students take tests to pass from one instructional module to another. If they fail one of these curricular embedded tests (C.E.T.'s), they study for awhile and take an equivalent form of the examination. When a certain percentage of students takes more than two equivalent tests, the unit is considered suspect and an analysis is made to improve its effectiveness. (23)

Plan tryouts so that as the data is collected, it is organized and readied for analysis.

One such tryout plan calls for recording student responses by asking them to press buttons or turn dials to indicate if they are learning, enjoying, and agreeing. The responses are transmitted by electrical impulse to a stylus which records the response on a sheet of paper. An individual or a summed group score can be recorded and scaled so that peaks and valleys of positive and negative reaction can be coordinated with instructional activity at a given moment. The system is attached to a computer which provides a numerical score. By the time a tryout is done, the summarized data is ready to be studied.

CASE 1

Producing Information Quickly

Audience Studies, Incorporated is a company which uses just such methods to test films, radio and television shows, and commercials. With their measures they can predict Nielsen ratings and box office returns. Sample audiences are recruited to answer questionnaires and fill out test forms and to allow measurement of physiological responses such as the basal skin response. Audience members are interviewed and taped. Staff members then ask the audience to respond on a dial which goes from very dull to very good during the presentation. Interest responses and the basal skin responses are automatically recorded as line charts by computer, so that in 24 hours a tryout report is ready for analysis. (24)

CASE 2

Producing Information Quickly

Staff members at the Southwest Regional Laboratory are preparing computer-controlled tryouts for any of the S.W.R.L. projects to make the constructive evaluation process at the lab easier. The computer unit consists of twelve tape recorders each of which will use an eight track tape of fifteen minutes duration. A total of 96 tapes can be stored in the machine. At present a twelve-button student response keyboard is planned for a student carrel, but later he will introduce a full keyboard and a light pen which would allow drawing. By pressing a code number a student may copy a lesson on a tape stored in the computer system. The tape is keyed to a video disk, similar to a large silver record, with 1760 tracks, each holding a still visual image which is duplicated on signal and monitored on one of the six television sets. The images can be labeled in different ways, and the visuals can be reviewed by a student.

A student will be able to sign on, name the tape desired, wait a minute for it to be copied, and then listen, watch and respond to the presentation. All responses will be automatically recorded, scored, and summarized. A student could be asked to view several units, thus reducing the cost as well.

The initial cost of $225,000 seems high, but, when S.W.R.L. staff members can provide many controlled lessons at once so easily, it seems worth the cost to them. Three full time staff members can put together the system, using a simple computer language. When the system functions, tryouts will essentially take care of themselves. (25)

Do not screen out interference; invite it. Know how the instructional system works under realistic and difficult conditions.

Prepare to test the system under toughest conditions when a good draft is available. But for early drafts, test them under relatively easy conditions to give the method a fair chance to show what it can do. (26)

The tryout should be designed to produce solutions as well as problems.

The tryout procedures and measures should at least reveal data which show strengths as well as weaknesses. At best, if you were to test a unit, you should get suggestions from students about ways to remedy the faults found in the unit.

A tryout should be considered a credible and trusted method by the people who make production decisions.

If you had someone helping you produce a unit -- a photographer and a writer, for example -- you should have them help plan and carry out the tryout.

CASE

Incorporating Production Staff in a Tryout

> Steve Klein of the Center for Evaluation at U.C.L.A. asks producers to accompany him to a tryout. The producer and he take turns as administrator of the tryout and as observer. Because of their trust in each other, and their trust in what is reported as hav[illegible] happened in the tryout, they can work cooperatively to make needed changes. (27)

More than one tryout and more than one test or observation in a tryout make the reporting more credible. Instead of one large tryout in one place, consider two smaller in-depth probes in two places.

> You should change your tryout approach or postpone a tryout when your resources are dwindling or are in question.

If your subjects, time, consensus on objectives, money for revision, subject matter, production time, or support of sponsors are reduced or changed in some way, consider a change in plans. You should also consider a change when your producer changes his attitude toward the evaluation plan or revision plan. You must maintain constant contact with a producer to detect this attitude change and to head off destructive changes. Be on the alert for data which reveals new evaluation questions.

You may want to make a change in the middle of a tryout. You may see a portion of a unit going very badly and you may want to ask everyone to leave that section alone. (28)

If an instructional segment is to be repeated or used for many students, or is directed toward a high priority objective, use most of the tryout criteria mentioned.

If a certain goal -- recognizing signs of malnutrition, for example -- is important to you, and your audience will be thousands of students, you must be rigorous.

Summary

To get the facts about an instructional program's strengths, you decide on a tryout plan including a combination of prototype units, tests, samples, and test sites. To be effective your tryout must be orderly, simple, properly timed, and based on valid data. You should conduct an economical tryout: one which uses few resources. Your tryout should be integrated into your program and should provide organized data ready for analysis. If a tryout is to help improve a program it must be naturalistic, it must be designed to produce solutions, and it must be a credible approach.

Organizing and Conducting a Tryout, in Brief

Choose a combination of prototypes, tests, samples, and test sites.

I AM GOING TO ADMINISTER THE INSTRUCTIONAL UNIT [AND TESTS] I HAVE SELECTED [choose any number from a. - d.]...

a. as is to a sample of students
b. and administer a comparable program to another sample of students [e.g., traditional version].
c. and administer tests but no program to another sample of students
d. and administer variations in the same unit to other samples of students [e.g., original and revised versions]

I AM GOING TO ADMINISTER A UNIT WHICH IS [choose one]...

a. a first draft
b. a rough but revised draft
c. a polished draft
d. a final draft

I AM GOING TO USE [choose any number from a. - g.]...

a. review
b. pre-test
c. post-test
d. progress test
e. rating form
f. interview
g. post-test for long term memory or application

MY TRYOUT WILL TAKE...

a. a short time
b. a long time

You continue planning by saying:

I WILL ADMINISTER THE UNIT AND TESTS TO A CERTAIN NUMBER OF SAMPLES OF [choose a combination]...

a. individuals
b. small groups (6 - 30)
c. large groups

THE PEOPLE (OR CLASSES) IN THE SAMPLE [choose a combination]...

a. are to be randomly assigned to the unit studied they are in [if more than one unit is used]
b. are to be randomly chosen from the target population
c. matched to other students in other groups based on certain characteristics [prerequisite abilities]

I WILL ADMINISTER MY PROGRAM AND TESTS TO MY SAMPLES IN A CERTAIN

NUMBER [choose a number] OF...
a. laboratory test sites
b. field test sites

A good tryout should be...

...orderly, constrained, and deliberate.

...based on valid data.

...able to yield statements of causation.

...easy, simple to remember and carry out, manageable and self-explanatory.

...a compromise among number of activities and amount of reliable information desired.

...run by a trained staff.

...adjusted to your production system.

...feasible within your resources, relatively inexpensive, acceptable to classroom teachers and require few subjects.

...fit into your instructional program.

...capable of providing organized data quickly.

...most realistic and complex.

...designed to produce solutions as well as problems.

...a credible and trusted method by those who make production decisions.

...changed when resources are in question.

...most rigorous when a segment is to be used many times for many students or meets a higher priority objective.

CHAPTER XVII

Assembling the Puzzle: Organizing the Data

To assemble a jigsaw puzzle, a person could put one piece into the puzzle at a time. Or, to make the job easier, he could organize his efforts by piecing together the border portions and pieces of similar color. The raw data, the answer, and the numbers collected from a tryout of an instructional project appear much like the jumble of jigsaw puzzle parts. To put the pieces of data together to get an accurate picture of what happened, you must organize your data.

The basic purpose for organizing data in constructive evaluation is that scoring, summarizing, and displaying data contribute to the improvement of instruction. You use this organized data to hunt for strengths and weaknesses in the program, and to answer evaluation questions. A good visual presentation of the data provides the essential picture of what happened.

Your scores can be based on comparison to some criterion, to objectives, or to a norm.

Numerical scores are normally the expressed results of achievement tests, but they could also be the expressed results for interview data. To do so, categories are first assigned to open-ended interview questions, then quantities are assigned to categories. It is relatively easy to label an answer and associate a quantity with it; the only problem in assigning such quantities is that the numbers may not mean anything.

CASE 1

Assigning Quantities to Free Response

The Dupont researchers assign numerical scores to answers given to the open-ended questions about recall on a commercial. For example, a -20 is scored when a subject

> "Can prove he was present during the time the commercial was aired but remembers nothing at all about the commercial."

A -50 is scored when...

> "A person who left the room during the commericial for a reason that did not demand his presence elsewhere. (A person who left to answer the door or the telephone or because the baby cried, etc. is not scored.)" (1)

CASE 2

Assigning Quantities to Free Response

When the Far West Laboratory asked parents 1) what they learned from the Toy Library training course that was useful and 2) what the most interesting part of the experience was, they had to develop a way to score the answers. The researchers reasoned,

> "In answering questions 1 and 2 the parents could:
>
> a. fail to respond, which was considered a negative reaction to the course;
> b. give a response they considered positive, but which was contrary to our objectives (for example, 'I learned to ask my child a lot of questions' or 'I learned it's good to make the child learn something every day') which was considered another negative response;
> c. give a response that was not contrary to but was not directly related to the objective, which was considered a neutral response;
> d. give a response that was related to the toys but was not directly related to the objective, which was considered a neutral response;
> e. give a response that related to the toys rather than to themselves or the child. This response was also considered neutral, because it indicated that the parents attributed the good things to the toys rather than to themselves;

f. give a response that was related to the objectives of the course. Furthermore, if the responses were positive and related to the objectives, they could be either so general that we could not relate them to a specific objective or they could be judged to be related to one of the objectives.

Therefore, we judged responses in this category to be:

(1) too general to classify;
(2) indicative of a feeling that the parents could help their children learn something useful;
(3) indicative of a feeling that the parents could influence the decisions that affect the education of their children; or
(4) indicative of a feeling that the child was capable or could be successful." (2)

The researchers also asked the parents what they didn't like about the course. The researchers organized responses into five categories because they felt the parent could...

"a. not respond (This was considered positive since they did respond to the first two questions);
b. make a positive response;
c. say nothing was wrong -- a positive response;
d. make a specific criticism;
e. be generally negative." (3)

CASE 3

Assigning Quantities to Achievement Data

For the purpose of detecting the precise strengths and weaknesses of a program, you could score by computing the number of students or proportion of students that reach an objective. Baldwin, (4) an educational researcher, showed a relatively simple method of scoring data by comparing results with a desired level of mastery. First, add up the scores for all students over the whole test, and divide that total by the number of items on the test multiplied by the number of students. For example, if three students scored 10, 20 and 30 on a 30-item test about sexuality, divide 60 points (10 + 20 + 30) by 90 (30 items x 3 students), and get .66 as the average mastery level for all students over all objectives measured on the test. If you were shooting for .80 and missed with a .66, Baldwin would suggest a more detailed score: add up the scores for all students on all items in a given category of objectives, and divide by the number of items in the category times the number of students. So, if you had 4 items on knowledge of sexual functions and 3 students scored 2, 3, and 4, divide 9 points (2 + 3 + 4) by 12 (4 items x 3 students) to get a .75 average level of mastery for all students on all knowledge of sexual function items.

To examine success in achieving individual objectives, simply divide the number of students who passed the item by the total number of students (15 students who passed divided by 20 who took the item) to get the proportion of students successfully completing the item (.75). Similar summaries can be used for feedback to individual students.

Scoring procedures should be reliable.

Use the same project staff member to score the same item through the test. You should check for consistency among raters, too.

To deliver the data to project producers quickly, use quick scoring techniques.

Students can be asked to self-score, or use automatic devices: chemical (special markers which make answers appear) or mechanical (computer). For multiple choice exams, a simple hole-punched answer sheet can be used as an overlay for quick scoring. Several companies have produced answer sheets upon which students indicate their response by erasing a square or marking with a felt tip pen. The erasure and mark reveal to the student the correctness of his answer. He may respond until he finds the correct answer, but the teacher knows if he first answered correctly and the test is scored by the student. Now a mimeograph device is available to make this sort of scoring possible for answers other than multiple choice.

Abedor's transparent overlay for his attitude scale is another useful technique. He tallied scores which showed negative attitudes for 30 questions on a transparency. When the five minute tally is finished Abedor proceeded to ask students to discuss their most negative feelings as shown by the highest number of tally marks.

Summarize your scores so you and your staff will be able to comprehend and use the results.

Summarizing is the method by which the data you have collected is simplified. You make your information as understandable as possible by using lists, tables, grids, graphs, and pictures. Some of the most commonly used displays are cumulative graphs, block frequency graphs, charts and tables.

You summarize your data in different ways for different purposes.

You may have several different evaluation questions or parts of questions which you want to answer, and so you summarize your results to answer each one in a fashion suitable for its nature. (5) (6)

CASE

Summarizing in Different Ways for Different Purposes

Roger Scott of the Southwest Regional Laboratory summarized and displayed the results of the Instructional Concepts Program in different ways for different purposes. (7) Some are shown below. As you may recall, children in kindergarten classes in several different schools were given the

Instructional Concepts Program and then tested to find out if they could identify examples of concepts by pointing. The data collected from these tests are summarized in the tables below.

The purpose of the summary in Table 1 is to answer questions about the similarities and differences there were among the eight groups taking the test.

TABLE 1

MEAN PRE-TEST AND POST-TEST CLASS SCORES
FOR THE CONCEPT IDENTIFICATION TEST

CLASS	NUMBER OF STUDENTS	MEAN NUMBER CORRECT		MEAN PERCENTAGE CORRECT	
		PRE-TEST	POST-TEST	PRE-TEST	POST-TEST
1	14	22.29	27.32	62%	76%
2	17	19.77	25.59	55%	71%
3	12	19.62	27.12	54%	75%
4	15	19.31	25.40	54%	70%
5	19	16.82	26.93	47%	75%
6	22	16.10	22.23	45%	62%
7	15	15.11	23.40	42%	65%
8	19	15.00	24.32	42%	67%
TOTAL	133	17.73	25.10	49%	70%

NOTE: All of the scores presented above have been corrected for guessing. The test contained 32 three-choice items and 4 two-choice items.

The purpose of the summary in Table 2 is to answer questions about the strengths and weaknesses in student performance on the concept test.

TABLE 2

MEAN PRE-TEST AND POST-TEST CONCEPT CATEGORY SCORES
FOR THE CONCEPT IDENTIFICATION TEST

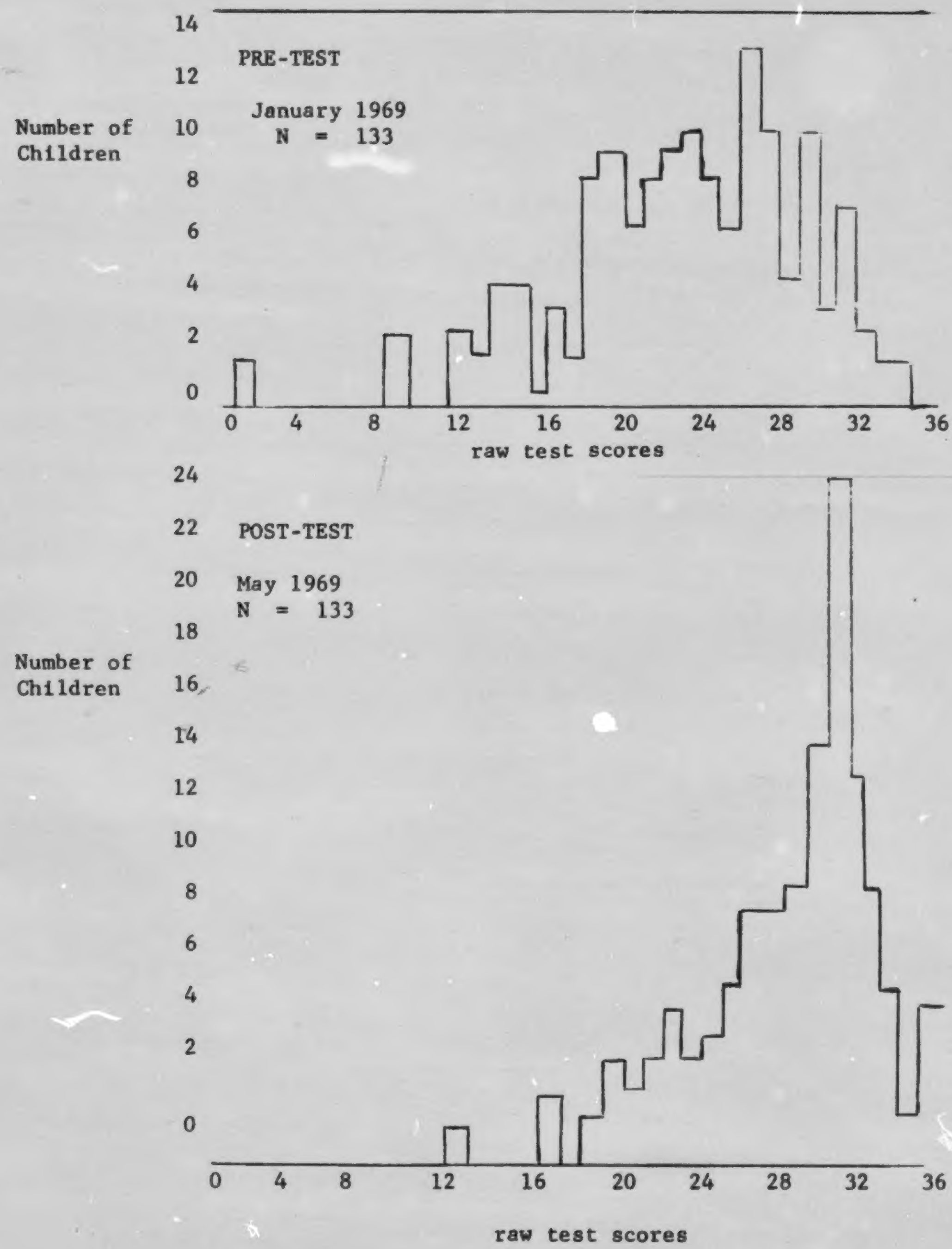
CONCEPT CATEGORY	MEAN NUMBER CORRECT		MEAN PERCENTAGE CORRECT	
	PRE-TEST	POST-TEST	PRE-TEST	POST-TEST
Color	3.98	4.78	80%	96%
Size	3.26	3.80	65%	76%
Conjunctive Concepts	2.75	3.98	55%	80%
Amount	2.41	3.20	48%	64%
Shape	1.84	3.34	37%	67%
Equivalence	1.82	2.71	36%	54%
Position	1.75	3.14	35%	63%
Time[a]	Less Than Chance	.15	Less Than Chance	15%
Total	17.73	25.10	49%	70%

NOTE: All of the scores presented above have been corrected for guessing.

[a] The Test contained only one item for the "Time" Category and five items each for the other seven categories.

The purpose of the summary in Figure 1 is to give a different view of the gain on the whole test. It shows graphically how a nearly random distribution turns into a skewed distribution when learning takes place. To make the comparison more dramatic Scott often superimposes one graph (using a dotted line) over the other (using a solid line).

Figure 1. Pre-test and post-test distributions for the concept identification test.



To answer questions about the comparison of gain scores of schools or the comparison of different concept categories Scott produced the following two tables. The examples shown are from an early tryout of the Instructional Concepts Program. (8)

TABLE 1

CONCEPT IDENTIFICATION TEST SCORES
FOR ELEVEN SCHOOLS

School	Mean Percent Correct		Percent Gain
	Pre-test	Post-test	
1	58.9	87.9	29.0
2	55.1	80.9	25.8
3	58.4	84.0	25.6
4	57.9	83.1	25.2
5	66.9	90.1	23.2
6	67.3	89.9	22.6
7	59.7	78.8	19.1
8	64.6	83.1	18.5
9	66.8	84.5	17.7
10	67.6	83.1	15.5
11	73.4	81.5	8.1

TABLE 2

SUBTEST SCORES FOR THE
CONCEPT IDENTIFICATION TEST

Subtest Concept Category	Mean Percent Correct		Percent Gain
	Pre-test	Post-test	
Colors	80.2	90.2	10.0
Shapes	48.9	82.0	33.1
Sizes	66.5	85.1	18.6
Positions	48.0	73.7	25.7
Amounts	53.3	82.2	28.9
Combinations	73.9	88.2	14.3
Comparisons	67.3	85.4	18.1
TOTAL	62.6	83.8	21.2

For the purpose of comprehending the answers to evaluation questions about the results of a teacher questionnaire on an early tryout of the Instructional Concepts Program, the Southwest Lab reported this way:

TABLE

TEACHER QUESTIONNAIRE
INSTRUCTIONAL CONCEPTS PROGRAM

Directions: Please give candid answers to the statements below. Do not sign your name.

Mark each item by circling one of the numbers as follows:
1 = strongly agree with statement
2 = agree
3 = neither agree nor disagree
4 = disagree
5 = strongly disagree

	Very much Agree				Very much Disagree
	1	2	3	4	5
1. With the program, I feel that my class is learning to identify instructional concepts.	24*	9	1	1	0
2. The program does not seem as useful to the children as the regular program used in our school.	0	2	6	12	13
3. The program takes too much classroom time.	2	5	5	11	12
4. The children participated eagerly in the program.	15	13	4	1	1
5. The teacher's manual did not provide sufficient guidance for me.	2	3	4	11	14
6. The lessons were too long for children of this age.	1	6	2	14	11
7. The program demanded too much time from the teacher.	1	5	4	15	9

*Written numbers refer to the frequency of each reported rating.

(9)

TEACHER QUESTIONNAIRE
(continued)

	Very much agree				Very much disagree
8. The children like the program as well as they like most activities at school.	1 ⑦ *	2 ⑱	3 ③	4 ①	5 ⓪

Mark each item by circling one of the numbers as follows:

1 = always true
2 = usually true
3 = sometimes true
4 = seldom true
5 = never true

	Always True				Never True
9. The children seemed to find the stories highly interesting.	1 ④	2 ⑫	3 ⑯	4 ②	5 ⓪
10. The objectives for each lesson were clear, worth-while goals.	1 ⑭	2 ⑬	3 ⑤	4 ②	5 ⓪
11. Materials were supplied to me in an easy-to-use form.	1 ㉑	2 ⑦	3 ⑥	4 ⓪	5 ⓪
12. My suggestions about the program were always well-received by SWRL representatives.	1 ㉘	2 ③	3 ①	4 ⓪	5 ⓪
13. The SWRL representatives who visited my class were very well informed about all aspects of the program.	1 ⑰	2 ⑬	3 ①	4 ①	5 ⓪

*Written numbers refer to the frequency of each reported rating.

(10)

TEACHER QUESTIONNAIRE
(page 3)

ACTIVITY	RATINGS					
STORY	12*	1	helpful to	20	1	easy to use
	12	2	children	4	2	
	5	3		7	3	
	4	4		2	4	
	0	5	not helpful	0	5	difficult
CONCEPT BOOKS	13	1	helpful to	18	1	easy to use
	12	2	children	8	2	
	8	3		2	3	
	1	4		2	4	
	0	5	not helpful	1	5	difficult
CONCEPT CARDS	15	1	helpful to	23	1	easy to use
	15	2	children	5	2	
	2	3		2	3	
	0	4		1	4	
	0	5	not helpful	0	5	difficult
GAMES	18	1	helpful to	12	1	easy to use
	9	2	children	6	2	
	3	3		11	3	
	1	4		2	4	
	1	5	not helpful	1	5	difficult
FLASHCARDS	21	1	helpful to	21	1	easy to use
	6	2	children	6	2	
	5	3		4	3	
	2	4		1	4	
	0	5	not helpful	0	5	difficult

THE ENTRIES BELOW ARE FOR THE LAST LESSON IN EACH UNIT ONLY.

PRACTICE EXERCISES	21	1	helpful to	23	1	easy to use
	10	2	children	6	2	
	2	3		3	3	
	2	4		0	4	
	0	5	not helpful	1	5	difficult
CRITERION EXERCISE	20	1	helpful to	22	1	easy to use
	8	2	children	5	2	
	5	3		3	3	
	2	4		1	4	
	0	5	not helpful	1	5	difficult

*Written numbers refer to the frequency of each reported rating.

(11)

To answer questions about which test scores are related to portions of the program you can use displays like a profile. Profiles have been used to pinpoint a particular problem during an instructional sequence. (12)

For the same purpose you can pin up scores on cards. Pin the cards next to test items on cards. Put these near sections of the program dealing with the topic. The score cards, test cards, and cards including sections of the program might be pinned to a wall or large bulletin board. You may discover that some preceding segment in a program contributes to the success or failure of a later one.

Summary

A good scoring and summarizing scheme provides a clear picture to serve as the basis for analysis: the puzzle is in one piece. To fully understand and be able to use the data you must use a method of analysis with a systematic and logical approach.

* * * * *

Organizing the Data, in Brief

Scores can be based on a comparison to...

...a criterion.

...a set of objectives.

...a norm.

Scoring procedures should be...

...reliable.

...efficient.

...comprehensive.

Data should be summarized differently for different purposes.

CHAPTER XVIII

Studying the Puzzle: Analyzing the Data

When a jigsaw puzzle is finally put together, a person can stand back, study the puzzle and say, "Now, let me see. How does the puzzle go together? Let me try to understand what the picture really means." In a similar fashion, you begin the analysis procedure for your constructive evaluation by studying your organized data. You list hypotheses explaining how instructional factors contributed to the results.

> First, you discriminate between standard and substandard results by checking scores against pre-established values.

You use the results from criterion tests, progress tests, attitude questionnaires, and interviews, and you compare them to cutoff points established for each standard and objective.

CASE

Distinguishing Substandard from Standard Results

Consider, for example, these exam and questionnaire results taken from a course package which proposed to teach students to apply psychological principles. On a post-class questionnaire, students were asked to say what they thought was the most difficult idea to learn in the class. Almost 50% of the students mentioned the idea "negative reinforcement." The students' statements brought the matter to the instructor's attention. To verify the problem reported, he checked the final exam items related to the concept "negative reinforcement." On four multiple choice items asking students to identify examples of the process of negative reinforcement -- putting someone in an uncomfortable situation from which he can escape -- only 60% of the 30 students chose the correct answers. On two questions calling for a written response applying the process of negative reinforcement to a case, only 66% answered correctly. The teacher considered these test results substandard because he considered the minimum passing score to be 80%.

To analyze efficiently the data you have collected in a tryout, you rank order the results, both standard and substandard by priorities.

CASE (continued)

Making Priorities Among Results

In the case of the psychology course, one of the most important objectives was for the students to learn to apply a list of principles. Among those principles was negative reinforcement. Because of the objective's importance, the instructor set out to find out the reason for the students' poor showing on test items relating to that principle.

To begin to detect the factors that contributed to positive and negative results, you study the details of the top priority test items and then try to infer which particular behavior made the performance good or poor.

CASE (continued)

Inferring Reasons for Results by Studying Test Items

When the instructor studied the wrong answers chosen by students on the multiple choice example, he found that most often students chose examples of punishment rather than examples of the process of negative reinforcement. When he studied the students' plans for application of the principle, he noticed the same trend. Students would prescribe punishment (transient unpleasant consequences) instead of negative reinforcement (a continual unpleasant state which one could escape by using the desired behavior). From this brief investigation, the instructor believed that the poor student performance was due to their inability to discriminate between punishment and the process of negative reinforcement. But he still had no firm hypothesis as to why they were unable to see the difference.

You must identify instructional segments which appear to relate to results.

You find and study instructional segments (units, chapters, paragraphs, slides portions of narration) which were created to contribute to priority objectives by using a prepared list of the segments which purport to influence certain objectives, information from student interviews concerning sections that they thought made learning contributions, and a task description that was used as the common basis for construction of tests and instructional segments. Or you may use evidence collected during the program tryout including progress tests, practice exercises, examples, outlines related to objectives, lesson plans, or indexed readings.

CASE (continued)

Finding Segments Related to Results

The psychology instructor collected course material and evidence which related to punishment and negative reinforcement principles. He found a chapter in the text, some course notes, handouts on the topic, a tape he had made, practice exam items, and student classroom assignments which were handed in. First, he observed that he had spent about 40 minutes of the tape time on punishment and followed with about 10 minutes on negative reinforcement. In addition, he had explained only the principle of negative reinforcement; its application had not been demonstrated. The two principles had been presented separately and not compared. He had given three relatively easy examples of negative reinforcement and students had been asked to respond to the examples by labeling, a performance unlike that required on the test. In practice exams and classroom assignments handed in, students had only a few instances to practice using the principle as it was required on the test.

> After finding instructional segments related to results and studying the connection, you are ready to infer the nature of the relationship between instructional method and results.

You state factors such as clarity of presentation or number of examples. To discover the factors, you can rely on theory, logic, or rules. If you had controls in your tryout you will be able to eliminate certain hypotheses.

CASE (continued)

Stating the Reason for Results

The psychology instructor believed that his students would learn to apply psychological principles if the course material highlighted critical attributes of principles, explained and demonstrated application of principles, and provided sufficient and appropriate practice in application of principles. Based on his belief and the evidence, he hypothesized that his students were unable to respond correctly to questions about negative reinforcement because he had not given sufficient explanation and demonstration in the taped lecture, he had not highlighted the critical attributes which differentiate punishment and negative reinforcement, nor had he compared them to other principles or other techniques. His examples of negative reinforcement were insufficient: he had given insufficient and inappropriate practice.

Because of resource limitations, you are not going to be able to test every hypothesis: often you have to choose among various hypotheses.

In choosing, you should consider these criteria:

a. the size of the tryout sample.

b. the adequacy of the criterion measures.

c. the consistency of the evidence.

d. the generality of the evidence.

e. the extent of the problem found.

f. the resources available.

CASE (continued)

Choosing Among Hypotheses to Test

The psychology instructor felt fairly certain of his hypothesis because it had been based on solid evidence: he had an adequate sample (30), the subtests and the questionnaire confirmed each other, and the problem seemed to affect a large percent of his students (about 40%). For him, the most convincing hypotheses about the students' failure to learn were 1) the confusing definition, 2) insufficient examples, and 3) insufficient practice. Less convincing were the hypotheses of 4) insufficient explanation and demonstration, and 5) highlighting of critical attributes.

The only question which remained was which of these hypotheses to act upon, given the resources available. The psychology instructor knew he had a limited but adequate budget for class handouts. Thus, he decided to give equal credibility to hypotheses three and four, and put hypothesis number one lowest in order of priority for action.

Summary

To find out what your data means, you analyze. You discriminate between standard and substandard results, and form priorities among results.

You study test items, identify instructional segments which appear to relate to results, and then infer the nature of the relationship. To be practical, you make priorities among hypotheses.

* * * * *

Analyzing the Data, in Brief

Distinguish between standard and substandard results.

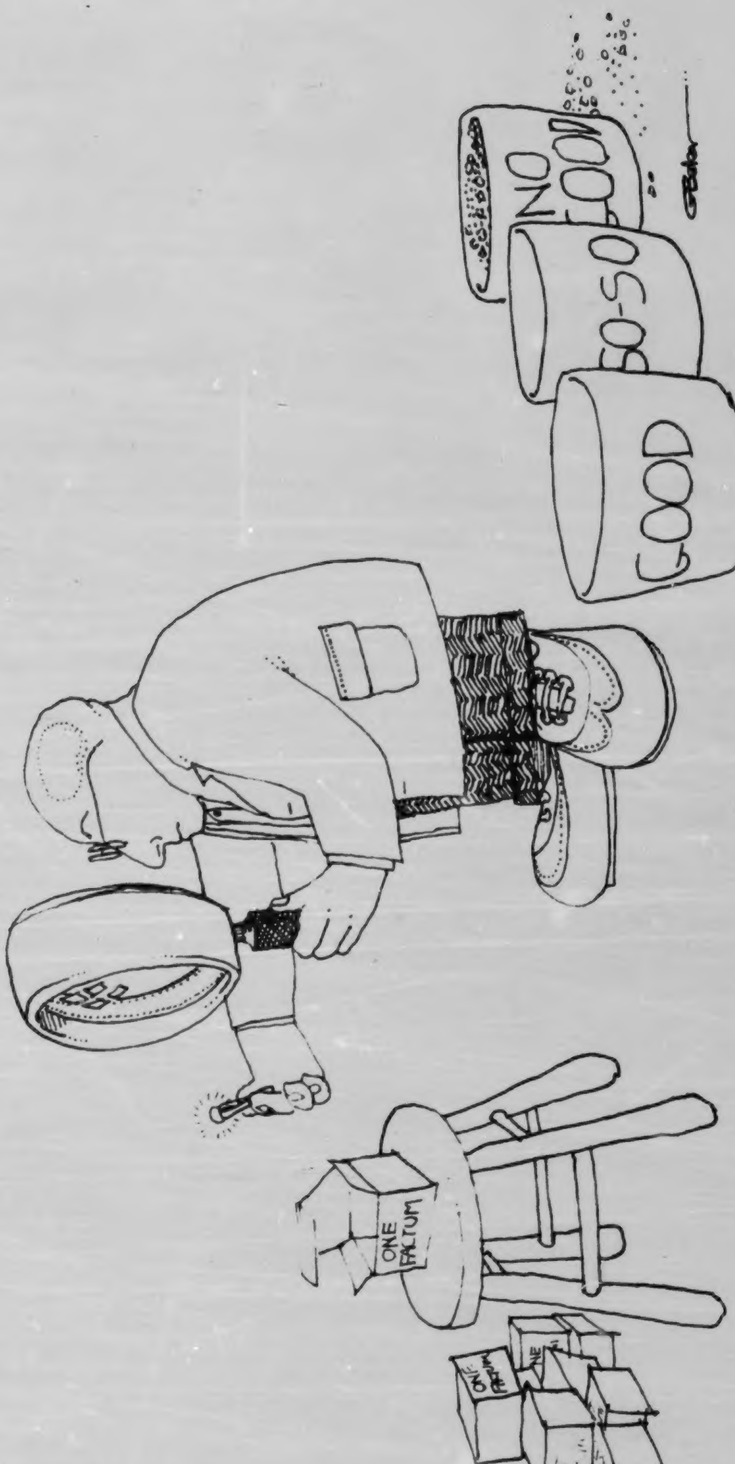
Make priorities among results to be analyzed.

Study top priority test items.

Identify instructional segments related to results.

Infer the reason for the test results.

Test the high priority hypotheses.


GOOD
SO-SO
NO GOOD
ONE FACTUM
ONE FACTUM
ONE FACTUM
ONE FACTUM

CHAPTER XIX

Detective at Work: Identifying the Strengths and Weaknesses of the Instructional Method

Before a detective can begin to ask "Who did it?", he must find out what was done and which of the events involved can be considered lawful acts and which could be called illegal acts. You must analyze the results of an instructional project in the same way to discover which results show success and which do not.

> In order to discover a program's strengths and weaknesses, you simply make a judgment about the adequacy of student response: does it fall within the boundaries of acceptable performance?

Before the evaluation takes place you set the cutoff points and you establish decision strategies. Set a limit for how many times an error in a program must appear before you will consider it a fault.

The comparison of result and standard is often an informal one: "Did we get 80% on objective three?" "No, we only got 75%." "Whoops, not good enough; we had better analyze that one and find out how to fix it." Or "We got 96%." "How did we do it?" "Let's find out so we can do it again."

To make a formal comparison you may want to use formulae to determine if a minimal level of mastery has been reached. If the student learning as computed by the formula reaches a certain level, you consider the program effective; if the calculated result falls below a limit set, the results are considered to indicate possible program faults.

There are many types of cutoffs which you can use:

1. Check the extent of the students' gain from pre-test to post-test.
2. Check the ratio of favorable to unfavorable responses on a questionnaire. (1)
3. Check the standard deviation. (2) (If the standard deviation is reduced from pre- to post-test then perhaps students who have low scores on the pre-test have reduced the gap between themselves and the rest of the group. You may infer that the groups were randomly distributed before and that perhaps now they are more alike.)
4. Look at averages.
5. Check high and low scores.

If students learned as much or more than was anticipated, the project can be considered successful; if, in addition, no negative side effects were found, the project can be considered even more successful. But if students did not meet the minimum standards for a segment, that part of the program is not considered a success.

CASE

Judging the Strength of a Unit

As mentioned before, researchers at the Individually Prescribed Instruction Project (3) check the number of mastery tests a student must take before he can show he has learned from the course materials: if a large number of students take more than two tests, an instructional designer checks the course materials. Here is an example of such a decision:

> "...The fact that such a small proportion of students shows mastery on the first CET (Curriculum Embedded Test) indicates that materials may be poor. Further study should attempt to determine whether this is true or whether poor performance is really due to such other factors as poor prescription (a poor method of teaching), invalid CET's, or a misplaced objective. Since these latter possibilities can be investigated by means discussed in other parts of this chapter, a complete study of the situation should be possible leading to an identification of the specific cause." (4)

TABLE 7

NUMBER OF PUPILS REQUIRING INDICATED NUMBER OF CET'S BEFORE SHOWING MASTERY OF OBJECTIVE

Objective 1		Objective 2		Objective 3	
No. of CET's	f*	No. of CET's	f	No. of CET's	f
1	3	1	29	1	8
2	16	2	9	2	28
3	15			3	2
4	4				

*f = students finishing the unit successfully

Sometimes the researchers convert the number of students taking more than one test to a proportion of the total number of students taking the unit. The proportion of .25 (no more than 25% of the students should be taking more than one post-test in a given unit) is designated as acceptable, but, when a percent excels .25, they check further. In the table below, for example, book three in unit one would be suspect, as would book three, unit eight.

A SEGMENT OF TABLE 35 (5)
PROPORTION OF PUPILS REQUIRING MORE THAN ONE POST-TEST IN SPECIFIED SPELLING BOOKS AND UNITS

Unit	Book Number		
	3	4	5
1	.43	.15	
2	.29	.06	.08
3	.39	.06	
4		.11	.05
5	.25	.30	.36
6	.09	.06	.16
7		.16	.29
8	.57	.11	

A comparison to a desired standard is still necessary even if you measure gain, improvement, or differences between groups. A change may be statistically significant but may not come close to a desired cutoff point. The basic question still stands: did the students achieve the program objectives? Therefore, you still must set a cutoff point and check the attainment of each objective.

> You may have to make further investigation to verify what seem to be existing strengths and weaknesses.

You may have to seek other sources of evidence or you may have to probe further into existing data. In the example which appeared earlier, the psychology professor interviewed some students about the problem he found which related to the principle of negative reinforcement.

CASE 1

Investigating to Verify a Strength

In the concepts program, Scott further analyzed existing data and found that, on the pre-test, students with Spanish surnames tended to score lower than students with non-Spanish surnames. The mean scores were 20.440 and 23.019. On the post-test, the scores for these groups were nearly identical -- 29.101 for Spanish surname students and 29.512 for non-Spanish surname students. (6) Hidden in the total score was an important program strength: the program was making a large difference to an important target population.

CASE 2

Investigating to Verify a Strength

To begin to answer questions about the validity of strengths and weaknesses in his genetics program, Anderson considered some of the seemingly unimportant data related to the program. He looked at the effect of variables not a part of the program: a student's entering behavior, the degree to which the program had been completed by a student, those students who follow practice instructions, and others.

He compared the average achievement scores of those who did study the program, with those who did not: 53.6% compared to 43.5% correct; not an earth-shattering result.

But when he accounted for other variables, he found certain strengths; for example, those students who never copied and who did finish the program attained an average of 70.5% of items on the achievement test. He summarized an estimate of the effect under optimal conditions according to each contributing factor. (7)

> To determine strengths and weaknesses accurately, base your decisions on test items which are related to objectives given to a large group of students in more than one well-planned tryout.

Evaluators often make judgments on a total test score. Remember: one person's score on a test reveals the same achievement as another's only if each item is related to the same objective, or if each objective is related to a specific set of items which are in order of difficulty.

Tests may indicate false gains or losses which are not related to the program, but collected scores from a large sample will average out influences outside the program, and precision tryout planning may prevent most false gains and losses which come about because of sloppy testing. (8)

No single test is conclusive. More than one item or test is necessary before you make major revisions based on supposed weaknesses. (9)

> You rank the strengths and weaknesses you find in order of priority based on the value of your objectives, the quality of your source of information, the degree of confirmation among results and the size of your audience.

The more important the goal, the more important the strength or weakness related to the result. The greater the importance of avoiding an error, the higher the priority of the strength or weakness. In curricula, errors related to an excess of content are less significant than errors of omission or commission.

Weigh findings based on student comments and criterion test results most heavily. In cases of technical judgment rely heavily and give priority to findings supported by technical reviews. (10)

Consider first those faults or strengths which have confirmation from several sources or several measures. Check to see if similar results are present in repeated tryouts. (11) Check to see if similar results are present in different instructional settings.

The larger your target audience for a given result, the higher the priority you place on that strength or fault. (12)

Summary

Comparing the results of a criterion test or an attitude questionnaire with a pre-established cutoff point, and pursuing that analysis further and setting priorities for results, is done to indicate what students have learned and where the program has succeeded or failed. It is preparation for a much more difficult task: figuring out the reason for the apparent successes and failures.

Identifying the Strengths and Weaknesses of an Instructional Method, in Brief

Ask if the results fall within acceptable boundaries.

Investigate further to verify strengths and weaknesses if necessary.

Put most of your confidence in results related to program objectives collected from many students in more than one tryout.

Order the strengths and weaknesses on the basis of

the value of the result,

the quality of your source of information,

the degree of confirmation among results, and

the size of the audience.

CHAPTER XX

The Puzzle of Keys and Locks: Identifying the Factors Which Contribute to Success and Failure

Imagine finding an old trunk with dozens of unusual keys, many locks, and a puzzling set of instructions: "One lock may need many keys; some locks only one. One key may fit many locks, some keys may fit none. Keys may unlock both locks and keys; if and when they please, locks may open locks, and on occasion, keys. Many keys have many locks, many more than in this box."

In any teaching situation there are many keys and locks. There are factors such as the nature of the presentation, the examples, and the practice; some factors contribute to attention, some contribute to motivation and learning. An educational situation is likely to be more complex than the puzzle of keys and locks. Single factors may have one effect (one type of example may make a concept clear); some may have several (a type of example may motivate, draw attention, and result in learning), and some factors may have no effect in a given situation. Some factors may contribute to others (several examples may constitute a presentation); some effects contribute to other effects (attention and motivation may contribute to learning). But few situations represent all factors and all results.

The job now is to sort out what factors you believe contribute to the results you have found. If you identify factors which contribute to both positive and negative results you will have a rational basis for revision: you can use your hypotheses to decide which factors to change to improve your results.

You may derive some plausible tentative hypotheses about the reasons for a program's results by studying the record of a student's responses.

You should inspect the records of your students' behavior as it was noted by observers or as it was recorded by the student in a test item. When a large number of students pass or fail a particular test item, for example, you can check the answer for a clue to the contributing factors. (1) But several students can pass or fail the same test item for different reasons. Only upon careful scrutiny can reasons be detected.

CASE

Inspecting Records of Student Test Performance

When Judy Light administered and tested a math program, she controlled many classroom conditions by standardizing them. When she found that a student did not pass all test items, both the test and the instruction were carefully analyzed to detect the fault.

How did she form her hypotheses? She asked herself five major questions; then she studied student responses on each test in which students did not pass every item. She studied the following example:

Write in the missing numbers using the associative principle.

(4x2)x5 = 4x(2x5)
= 4x 10
= 40

(9x3)x6 = 9x(3x6)
X = 27 x 18
= ___

2x(4x8) = (2x4)x 8
X = 32 x 8
= 256

6x(7x4) = (6x7)x 4
X = 28 x 42
= ___

The first two questions were:

"1. What was similar about the problems missed on the test?
    a. The student always made the first error on the second line of the problem.
    b. The errors appear to be systematic. The pupil always puts the product of the multiplication problems within both sets of parentheses from the first line into the blanks on the second line.
2. How did the items missed differ from those items passed on the test?
    a. The one item passed had one numeral, a 4, already written in the second line." (2)

After she studied test responses, she related her observations to program materials. Judy Light reasoned at this point that perhaps the student had not learned the associative principle but the materials seemed to have clearly explained the rule. Finally, her attention focused on the page before the test. This is the last page before the test.

Multiplication is associative:

(8x2)x2 = 8x(2x2)
↓ ↓
16 x2 = 8x 4

32 = 32

Write in the missing numbers and solve the equation using the associative principle:

(3x2)x5 = 3x(2x 5)
6 x5 = 3x 10
___ = ___

(3x9)x4 = 3x(___x4)
↓ ↓
___ x4 = 3x ___
___ = ___

(7x 6)x3 = 7x(6x3)
↓
___ x3 = 7x___
___ = ___

Judy Light continued to ask:

"3. Where in materials were items presented?
   a. The format on this page differed from the test. The student was always required to write in the product of the multiplication problems within the parentheses in the second line.
   b. The student also always had an arrow to aid him in putting the product in the correct place.
   c. This page also differed from the test in that the student solved each problem for both equation types (axb)xc and ax(bxc). On the test he was required to solve only one side of the equation, eliminating a check of his work." (3)

Once sufficient evidence had been gathered, Judy Light made a hypothesis and decided on a revision.

4. What caused the failure?
   Hypothesis to be tested:
   If the last page of the materials is changed to include problems similar to the test, then the student will pass the test.
5. How can the hypothesized cause of failure be tested?
   The following page was added as the last page in the materials. The student does not have arrows to indicate where the products are placed and he only answers one side of the equation.

Solve each equation:

(2x5)x3 = 2x(5x3)
= ___x___
= ____

(3x1)x2 = 3x(1x2)
= ___x___
= ____

(2x7)x3 = 2x(7x3)
= ___x___
= ____

(8x1)x3 = 8x(1x3)
= ___x___
= ____

(3x5)x6 = 3x(5x6)
= ___x___
= ____

This was a relatively simple analysis; many more complex analyses illustrate her work. Complex analyses of this sort requires a subject matter expert: one who can state all the steps and decisions of a task and all the prerequisite knowledge required.

Light's results showed that student performance improved on 82% of the objectives analyzed. Of 55 objectives, students reached the criterion level on 27, improved on 18, remained the same on 7, and did worse on only 3.

You should consult prepared aids which link instruction and results.

You locate instructional segments (chapters, paragraphs, examples, and practice exercises) associated with priority objectives, and deal first with those objectives not achieved. For this purpose you may use several kinds of prepared aids. It is best if these aids are prepared before analysis begins, but there is no reason why they cannot be created as they are needed.

Table: Aids for Linking Instruction and Results

TYPE OF AID	DESCRIPTION OF AID	HOW TO USE AID
Knowledge structures and diagrams	Lists or diagrams are made of the structure of a discipline or subject, relating its ideas.	Look for ideas which may be related to the results you are analyzing.
Lists of relationships	A plan is drawn relating instructional factors (presentation variables and practice variables) to likely results.	For the results you are considering, look for related factors in the materials.
Task descriptions	Each step and decision of a given task is outlined or diagrammed. A list is made of each concept, task, skill, or principle which is prerequisite for a task to be learned.	Look for instructional material related to the steps, decisions, or prerequisite knowledge of the result under study.
Outlines and indexes	As in textbooks, an index is made which shows where topics are handled.	Find all references to the content of the result being studied.
Lesson plans	Specific instructional activities are related to objectives.	Look up the activities related to the result being investigated.

CASE

Using Prepared Aids

The psychology teacher who found that students ran into a problem learning about negative reinforcement used prepared aids. He used lesson plans which helped him find the instructional activities related to the results he was investigating. His lesson plan outlines looked like this:

1. Students listen to lecture, including definition and examples.
2. Students practice examples in class.
3. Students do practice tests at home.
4. Students read text at home.

This led him to look at the lectures, classroom practice, practice tests, and the text for their possible contributions to the result.

He used the index in the textbook he assigned in order to find all references to the principle that students could read. He had an index of his course notes and handouts like this:

negative reinforcement:

definition	p. 53
examples	pp. 53, 54
practice items test	p. 65
appendix	pp. 12, 15, 20

This saved considerable time by helping the instructor find all the practice test items which did contribute to the errors the students were making.

There are three other aids he used which provided a link and tentative hypotheses, too. He used a list or diagram like the one following which relates ideas in the psychology of behavior to each other:

Negative Reinforcement is defined as

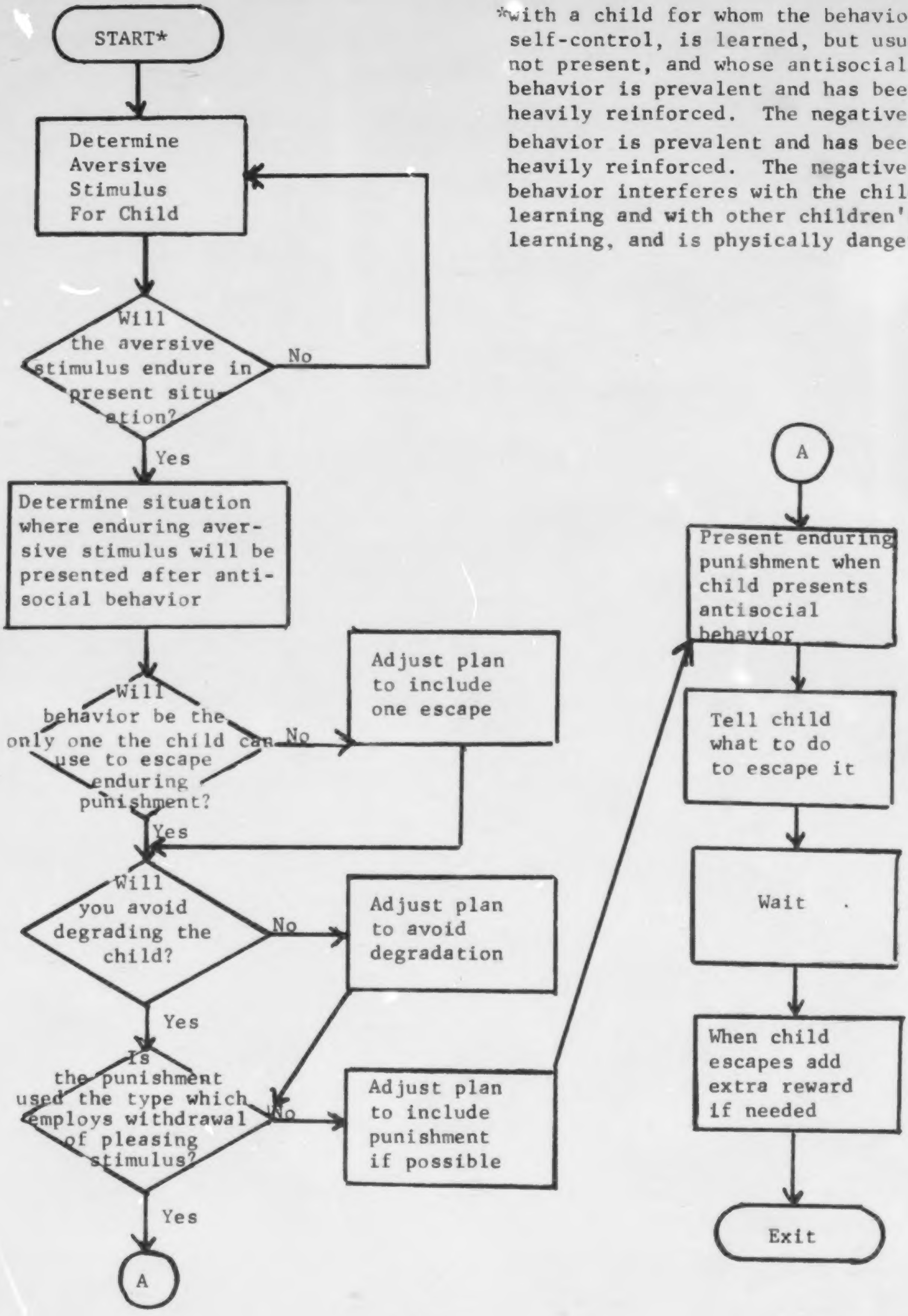
punishment 1 present unpleasant stimulus	a) enduring punishment	b) followed by a behavior which results in	c) withdrawing of the aversive stimulus
punishment 2	not		not
withdraw unpleasant stimulus	transient punishment		adding a pleasing stimulus, e.g., novel primary or secondary reinforcers

The diagram led him to check to see if he had stressed and if the students knew each of the three major parts of the definition, including both forms of punishment as contributors to the process of negative reinforcement, and distinguishing between enduring and transient punishment, and withdrawing an aversive stimulus and adding a pleasant one. The absence or distortion of any of these characteristics of the principle would be cause to believe that the gap or error was a factor contributing to the result.

He listed the relationships of instructional factors and results he planned to use to get students to reach the objectives of the program. His list included the following:

1. Provide practice in prerequisite knowledge and skill to enable students to learn and apply the principle.
2. Provide direct practice to facilitate the immediate transfer to apply principle.
3. Provide a precise definition highlighting each attribute to enable a student to recognize the principle in use.
4. Provide many diverse examples of the use of the principle to enable students to apply to many situations.
5. Provide a demonstration of the application so that students can imitate it.
6. Provide several different demonstrations so that students can generalize about it.

He used the list as a series of checkpoints: reviewing the material and checking for the presence and correct use of each factor. If the factor was not present or was incorrectly used, he considered the factor as contributing to the result. Simultaneously, he studied a task description.


START*
Determine Aversive Stimulus For Child
Will the aversive stimulus endure in present situation?
No
Yes
Determine situation where enduring aversive stimulus will be presented after antisocial behavior
Will behavior be the only one the child can use to escape enduring punishment?
No
Adjust plan to include one escape
Yes
Will you avoid degrading the child?
No
Adjust plan to avoid degradation
Yes
Is the punishment used the type which employs withdrawal of pleasing stimulus?
No
Adjust plan to include punishment if possible
Yes
A
A
Present enduring punishment when child presents antisocial behavior
Tell child what to do to escape it
Wait
When child escapes add extra reward if needed
Exit

*with a child for whom the behavior, self-control, is learned, but usually not present, and whose antisocial behavior is prevalent and has been heavily reinforced. The negative behavior is prevalent and has been heavily reinforced. The negative behavior interferes with the child's learning and with other children's learning, and is physically dangerous.

The psychology teacher analyzed the task description for all the steps, decisions, and knowledge required to do the task as described. He checked to see if there was instructional material, both presentation and practice, which could help the student to perform the task as described.

You should consider the data which show a relationship between portions of an instructional program and results.

TABLE Description and uses of data sources about factors contributing to instructional effectiveness

Data Source	Description of Data Source	Use of Data Source
Interview	Transcripts or notes from discussion about the aspects of instruction which seemed to make it work or fail.	To extract subjective impressions as to what factors contributed at what time.
Practice Exams	Scores on practice exams or exercises taken during the course of instruction.	To find clues as to where students began to make errors and the sort of errors they first made.
Progress Tests	Scores on measures of behavior associated with learning during instruction.	To find behaviors which may influence learning and their relation to segments of the course.
Diagnostic Tests	Scores from tests constructed to generate hypotheses about contributing factors.	Then study results for a link to the portion of the material related to the results or type of error made.

Transcripts or notes of student interviews provide hypotheses.

Sometimes the students generate some insightful notions about why the program did well or poorly. You can help students generate hypotheses by encouraging them in an open interview.

CASE 1

Using An Interview to Generate Hypotheses

Abedor was able to encourage students to state meaningful hypotheses during his group interview process. Abedor reports

> "...students indicated that the post-test was unfair, in that it was not a representative sample of lesson content. This, in spite of the fact that E [the experimenter, Abedor] and Author A had agreed that the post-test adequately sampled student knowledge with respect to the lesson objectives. After some discussion, it became clear that the problem did not lie in the post-test, which did, in fact, test lesson objectives. The problem was in the relative emphasis given certain content in the SLATE--which was not reflected either in the lesson objectives or the post-test. Specifically, 15 minutes of one SLATE were spent on historical development of the cattle industry (with numerous places, dates, and other historical information). Knowledge of historical development was not a major objective of the lesson, consequently only two (out of fifty) post-test items referred to historical development. The students, in the meantime, had been concentrating on memorizing the historical part at the expense of the other concepts. The debriefing, therefore, had explicated the combination of factors which led to this feeling of frustration on the part of students; namely, they didn't read the objectives, and the SLATE content overemphasized that which was not a lesson objective." (4)

CASE 2

Using An Interview in Advertising Research

Interviews are often used as a source of hypotheses in constructive evaluation in advertising. In the first television ad produced for No More Tangles Shampoo, a mother and a young daughter were shown demonstrating the product's virtues. The mother explained the hair snags and tangles would be gone from the little girl's hair if the product was used. When questioned, many of the women in the viewing audience said that the product solves a child's problem. This was enough of a clue for the writers to make a production change. In the revised commercial, the camera focused on a long-haired five-year-old girl who explained how No More Tangles solved her hair problem. The scores on learning and attitude measures used leaped an average of twenty percent. (5)

> If you can find records of a student's performance on practice exams, you may be able to find out where the student began to make errors.

You can study the type of errors you find and derive some hypotheses.

CASE

Using Records of Practice Exams

Judy Light's analysis of the last practice page in the math program was an example of this procedure. As you read in an earlier example, she found out precisely where the student made an error related to the exam requirements and then analyzed the student's mistake and the instruction associated with it. (6)

> From progress tests, you may derive hypotheses which describe the influence of these behaviors on a student's criterion performance.

To find contributing factors you must discover precisely what happened during instruction. For example, if during instruction a student fails to complete a response, or if a student hesitates, acts bored, does not follow proper sequence, does not attend, you may have a clue as to why he did not learn.

CASE

Using Data From Progress Tests

If children failed to learn a sight word from "The Electric Company," researchers could hypothesize that the children's attention lagged at the points in the program which the word was shown. They might look at distractor data to validate their hypothesis. If they found a low attention score related to the sight word, they would explore the program segments to find factors common to the segments which killed attention.

It might be that every time the sight word was presented, excessive dialogue was used, or the student didn't understand the premise of the segment, or that an actor's movement distracted the student from the word. Other data may be checked or collected to validate any one of these hypotheses. (7)

You may use diagnostic tests to generate hypotheses, tests consisting of items which ask for all the knowledge and skill related to a final requirement.

The data from the diagnostic test shows precisely which subskills the student has not learned because the items are directly related to portions of the course material.

CASE 1

Using a Diagnostic Test

Fitzpatrick, an instructional developer, created a course on economic analysis which used this technique. He states:

> "The way in which learning packages were constructed made it possible to identify with great precision where a change had to be made in a segment

to improve it. In the Self-Instructional Printed Packages, for example, an analysis of performance on the criterion test would include the segment, page within the segment, paragraph on the page, and sentence in the paragraph that caused the learning difficulty." (8)

CASE 2

Using a Diagnostic Test

The test items can be constructed so that the answers reveal the instructional problems, providing considerable help in analyzing test performance. For example, the psychology instructor who had some difficulty in teaching negative reinforcement could ask, "Which of the following is an example of negative reinforcement?"

(a) a teacher keeps a child in the hall until she thinks he's ready to come out

(b) a teacher puts a child in the hall and asks him to come back when he finishes his assignment

(c) a teacher puts a child in the hall

(d) a teacher spanks a child and asks him to finish the assignment

(e) a teacher spanks the child

The process of negative reinforcement is the presentation of enduring punishment with the possibility of escape by manifesting the desired behavior. Thus, the correct choice would have to be (b). If a student chose (a), he may have not known about the attribute "escape." If he chosen (c), he may be unaware of the notion of escape and equates negative reinforcement with one form of punishment. If he chose (d), he may not be aware of the attribute of enduring vs. transient punishment. If he chose (e), he may be equating the principle with one of the forms of punishment.

It may be possible to find a certain program segment linked to a certain result--but to find the reason for the link, you will need some guiding theory.

To make inferences, you consider your results and you think of principles which predict similar results. If students are not paying attention, for example, you think of principles which include attention. Check to see if any factors noted in the principle are present in the evidence you have collected. If you discover three factors contributing to attention--for example, novelty, reward, and meaningfulness--you check for the presence or absence of these factors in your program. From this exploration, you may discover that the most likely contributing factor is novelty. Your hypothesis would be, for example, the repetition or lack of novelty contributed to the students' lack of attention and subsequent failure to learn.

> There are many heuristics, operating procedures, or rules-of-thumb which can help you form hypotheses.

Heuristics bear close resemblance to theoretical principles, but do not have strong empirical support. They are usually derived from personal experiences, case studies, and informal research studies.

CASE 1

Using Rules-of-Thumb

Ken O'Bryan, of The Ontario Institute for Studies in Education, made eye movement measurements of a number of good readers, slow readers, and non-readers watching "The Electric Company." During the summer of 1972, Ken O'Bryan stated his first general impression about his findings. Because of the relatively tentative nature of the results, his statements can be considered rules-of-thumb or heuristics. At the time of this writing (Summer, 1973) O'Bryan has replicated his results with more children and confirmed his early statements. Given this added evidence, the generalizations begin to border on empirically supported principles.

Excerpt from Memo on Eye-Movement, July 31, 1972

"The general, (with regard to differences between groups of children,) findings were as follows:

1. All good readers (Group A) showed normal reading patterns, also exhibited by adult readers.

2. Slow readers (Group B) are somewhat slower to orient to new material, and are more easily distracted by action and by speaker's face. This group requires more time to fixate on the material, and when interrupted in this process, will start over at the beginning of the word. In general, the poor reader exhibits the same eye-movement patterns as the good reader, but at a much slower pace. This is an important production situation.

3. Non-readers (Group C) exhibit largely random eye-movements. The print is given little systematic attention. They are drawn strongly to action, and are extremely slow to orient to new material as it appears on the screen. However, this group did tend to fixate longer on flashing letters than the other children.

The following general findings apply to the bits themselves, rather than to differences between children's reading levels:

1. Whenever talking occurs (as in 'Row, Row, Row Your Boat'), children tend to look away from the print and at the speaker. The poor reader, when thus interrupted, is forced to start over again, and often never finishes reading the word or phrase.

2. Action in animated bits is less distracting than animation in live bits; perhaps because it is uncluttered.

3. Animating the word itself is highly successful in producing left-to-right scanning by the child.

4. In general, when the word carries the action (is on a character's shirt, or an important prop, for example) focus on the word by the child is good.

5. Eye-movements in repetitive segments (like 'The Surgeon') do not show that the child looks more at the word once he has 'had his fill' of the character, as might be expected. Eye-movements are essentially the same (dwelling on the face) throughout the bit.

6. Print is best presented in a central position, at eye or mouth level. The lower part of the screen is the worst place to put the print. Near the top of the screen is slightly better." (9)

CASE 2

Using Heuristics

Judy Light presents a number of heuristics which are useful in deriving hypotheses about reasons for program success or failure. Some are phrased as "if...then" statements: (10)

If a pupil fails a curriculum-embedded test, then

a. the pages may not teach and provide practice on the tested content.
b. the pages may not teach and provide practice on "unique" properties.
c. the pages may not require adequate practice.
d. the prescription may not contain pages which are duplicates in form and content of the CET (curriculum-embedded test)
e. the prescription may be inadequate.
f. the pages may not provide practice involving the same format as the test.
g. he (the student) may not have learned from the teaching pages.
h. his work may demonstrate poor work skills.
i. he may have done the prescription incorrectly.
j. he may not have the appropriate prerequisite behaviors.
k. he may not be motivated to do accurate work.
l. he may not be "attending to task" while doing his work.
m. he may not be checking his work.
n. he may not be able to use self-evaluation skills to decide if he has learned the required skills.

If a pupil has failed an objective on the post-test (11)

a. and passed the objective on the pre-test, then the pre-test and post-test may not be parallel forms.
b. and passed the CET, then the CET and post-test may not be equivalent in either form or content.
c. and passed the CET, then the prescription may not provide enough practice for learning to occur.
d. and passed the CET, then the pages and CET may not teach him how to discriminate directions.
e. and passed the CET, then he may not have sufficiently reviewed before taking the test.
f. and passed the CET, then he may not have checked over his work.
g. and passed the CET, then the criterion for mastery performance may not be adequate.
h. then he may not be motivated to pass the test.
i. then he may not have been "attending to task" while taking the test.

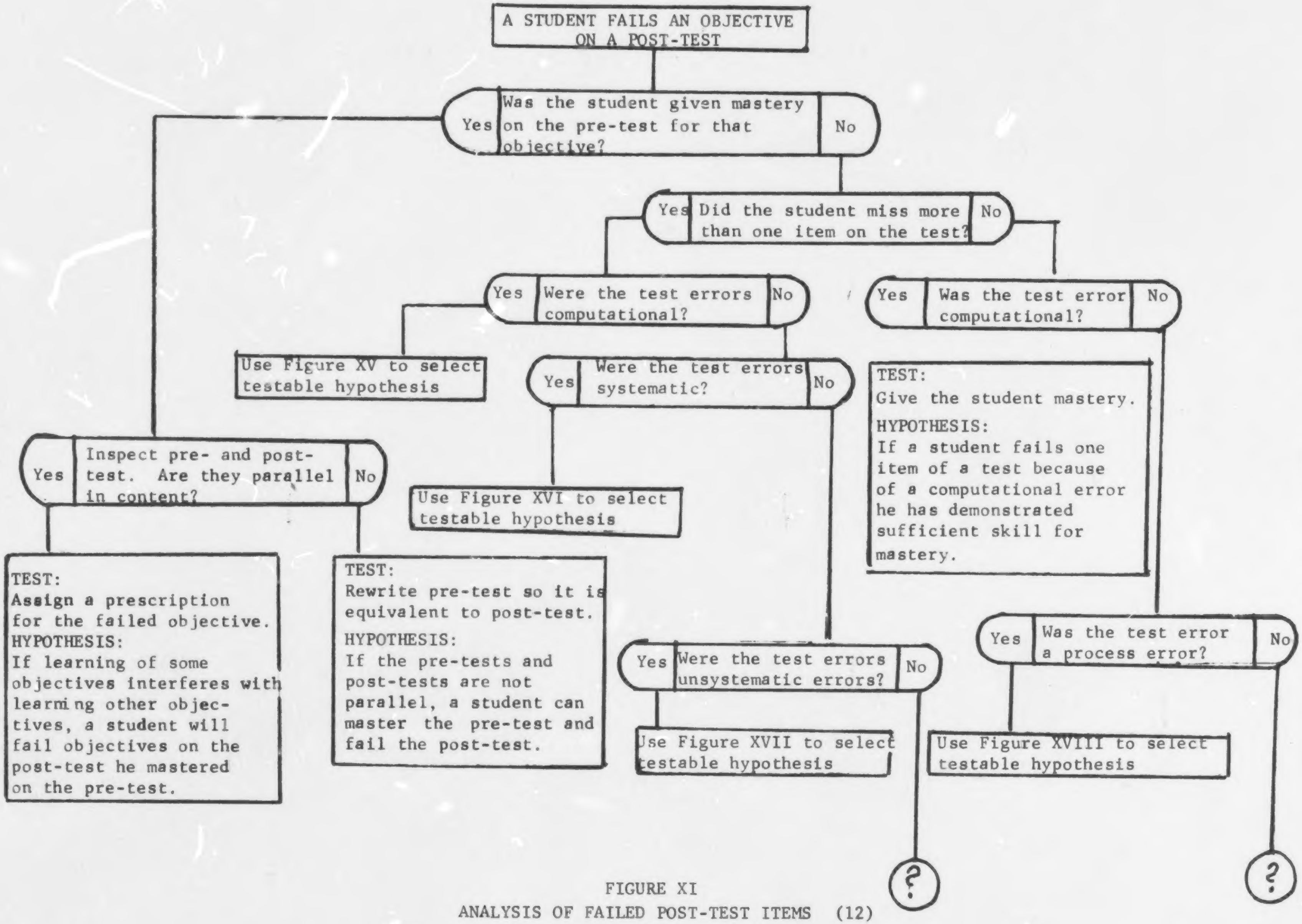

A STUDENT FAILS AN OBJECTIVE ON A POST-TEST
Was the student given mastery on the pre-test for that objective?
Yes
No
Did the student miss more than one item on the test?
Yes
No
Were the test errors computational?
Yes
No
Use Figure XV to select testable hypothesis
Were the test errors systematic?
Yes
No
Use Figure XVI to select testable hypothesis
Were the test errors unsystematic errors?
Yes
No
Use Figure XVII to select testable hypothesis
?
Was the test error computational?
Yes
No
TEST:
Give the student mastery.
HYPOTHESIS:
If a student fails one item of a test because of a computational error he has demonstrated sufficient skill for mastery.
Was the test error a process error?
Yes
No
Use Figure XVIII to select testable hypothesis
?
Inspect pre- and post-test. Are they parallel in content?
Yes
No
TEST:
Assign a prescription for the failed objective.
HYPOTHESIS:
If learning of some objectives interferes with learning other objectives, a student will fail objectives on the post-test he mastered on the pre-test.
TEST:
Rewrite pre-test so it is equivalent to post-test.
HYPOTHESIS:
If the pre-tests and post-tests are not parallel, a student can master the pre-test and fail the post-test.

FIGURE XI
ANALYSIS OF FAILED POST-TEST ITEMS (12)

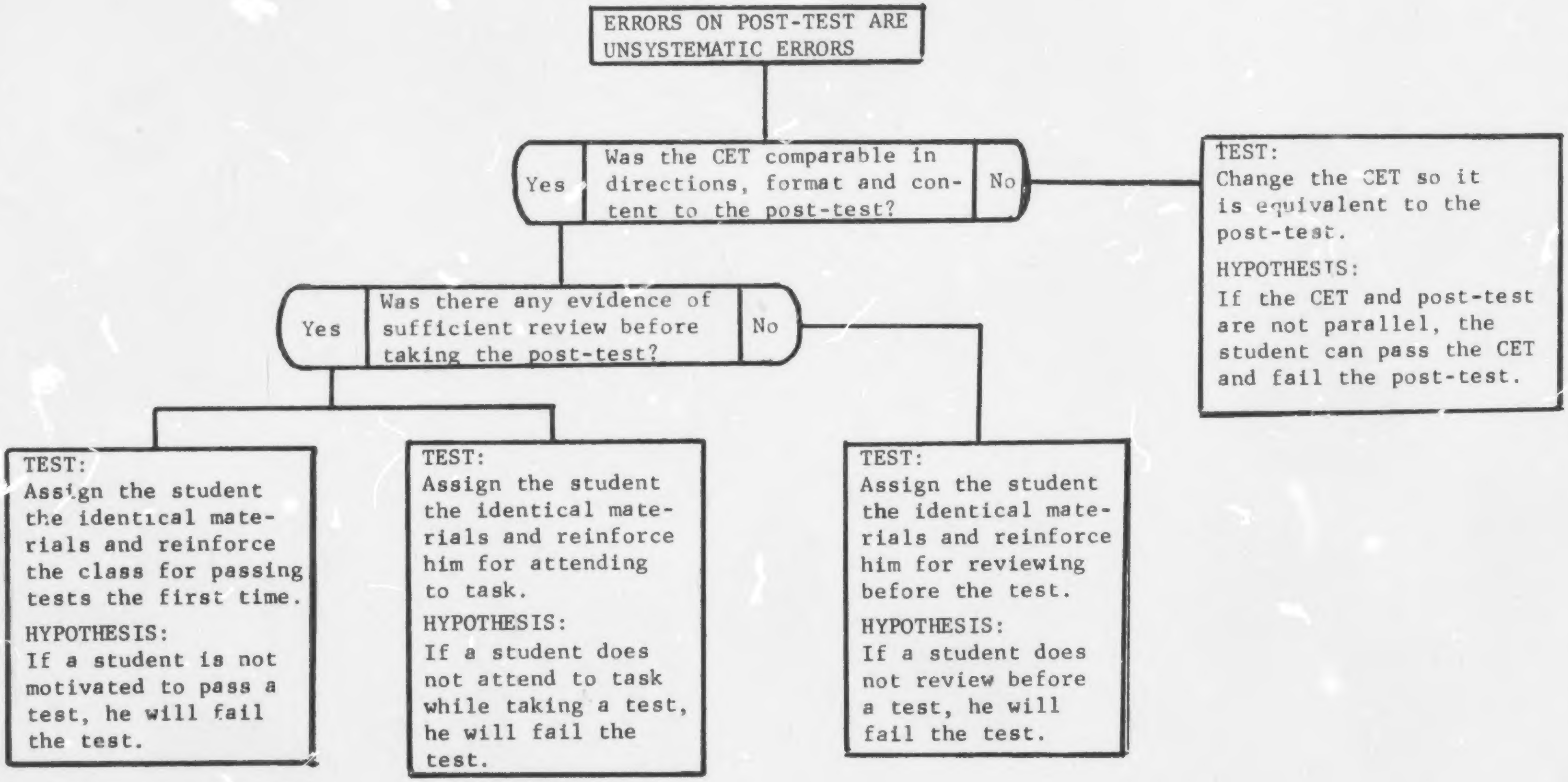

ERRORS ON POST-TEST ARE UNSYSTEMATIC ERRORS
Yes
Was the CET comparable in directions, format and content to the post-test?
No
TEST:
Change the CET so it is equivalent to the post-test.
HYPOTHESIS:
If the CET and post-test are not parallel, the student can pass the CET and fail the post-test.
Yes
Was there any evidence of sufficient review before taking the post-test?
No
TEST:
Assign the student the identical materials and reinforce the class for passing tests the first time.
HYPOTHESIS:
If a student is not motivated to pass a test, he will fail the test.
TEST:
Assign the student the identical materials and reinforce him for attending to task.
HYPOTHESIS:
If a student does not attend to task while taking a test, he will fail the test.
TEST:
Assign the student the identical materials and reinforce him for reviewing before the test.
HYPOTHESIS:
If a student does not review before a test, he will fail the test.

FIGURE XVII

ANALYSIS OF UNSYSTEMATIC ERRORS (13)

Producers will not follow a set of instructions which tells them which sources of evidence to use and how to use them.

Your producer should form the hypotheses about factors contributing to success and failure in his own way, because those who collect and summarize the data should not be the ones who draw inferences from it. But your producer should be encouraged to make his hypotheses on several sources of evidence. He should choose a few, important, credible data sources to summarize and integrate all the complex data sources.

CASE

Resisting Instructions to Use Sources

Dick, an instructional researcher, used a checklist of seven sources of feedback to help inexperienced programmers revise a program. He gave them post-test item analyses, error rates, student comments, teacher comments, correct and incorrect answers for all items, and a page number where ideas for each item were taught in the text. (14)

He gave them a handout including these instructions:

1. Study the item analysis of the end-of-lesson test to determine those concepts which were most often missed by the students.

2. Study the incorrect responses to these particular test items to determine if there was a straightforward misunderstanding of notation, a complete lack of comprehension of the concept, or a variety of errors.

3. Use the guide to determine those frames in the program which dealt most directly with the concept(s) missed on the test.

4. Study the student error rates for these frames. If the program frames are quite similar to the test items and the error rate is quite low, more practice frames should be provided. If the error rate is quite high, these frames need revision.

5. Study the sample of incorrect student responses to this segment of the program. These responses should suggest the nature of the learning difficulty and the type of revision needed.

6. Study the comments of both the students and the program reviewers for further suggestions concerning the problems encountered with these particular frames.

7. If no frames in the program correspond to a test item missed by a large percentage of the students, consider the addition of frames that will "bridge the gap" between the present learning materials and what would be considered a transfer type item.

The programmers used the information on error rate and teacher comments to make their decisions. If the student error rate was large, then they checked student comments. None of them followed the rules as they were stated, and few used the item analyses and the test items related to text pages.

The programmers complained that the test (which they had not constructed) did not measure the objectives, and they stated that they wanted to know the level of ability of students making comments. The programmers preferred summarized data from many students rather than detailed information from single students.

> You may need to collect more data because it may be that, even with aids, data, theory, and rules, you could still be puzzled about what contributed to success and failure.

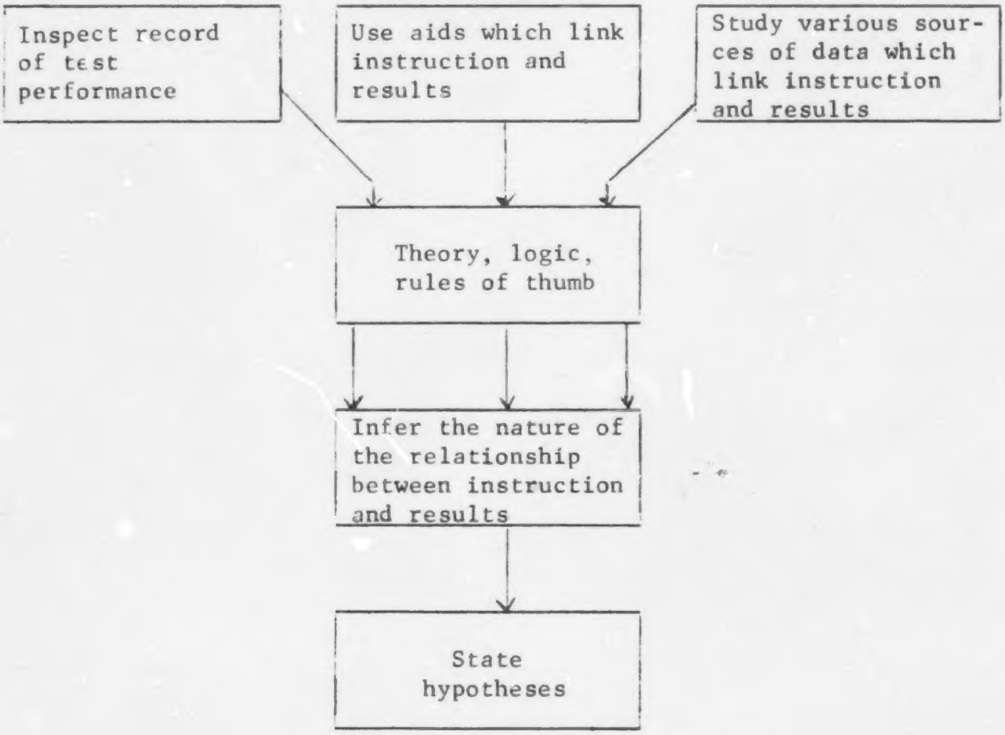
You may have some tentative hypotheses related to rules and theory: you may have hypotheses which you have discovered yourself (because there are no theories or rules), or you may have no hypotheses at all. At this stage, in many cases, project directors find or confirm hypotheses through further testing. (15)

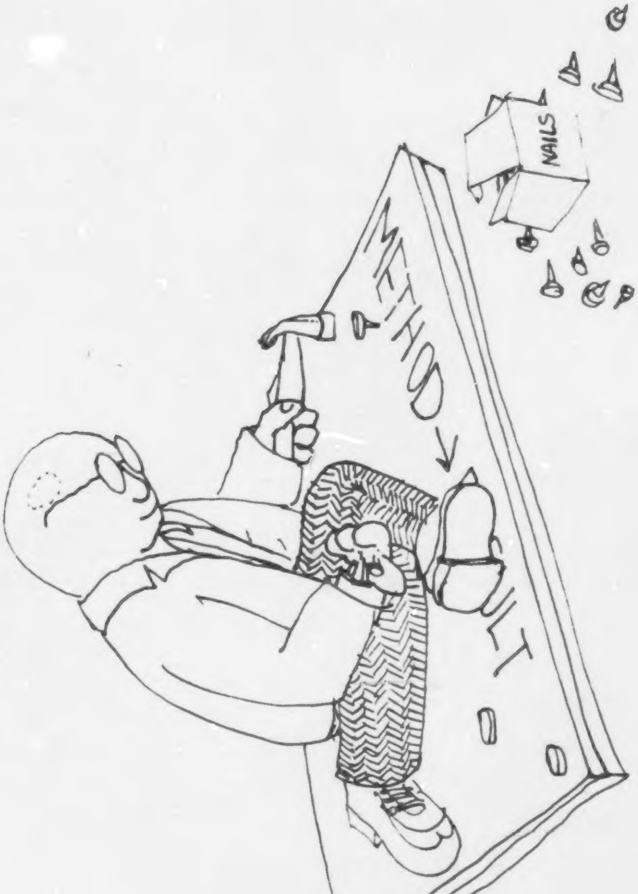
Summary

There are three major sources of evidence which you can use to make hypotheses: records of test performance, aids and data which link instruction and results. The tentative hypotheses are filtered through a sieve of theory, logic, or heuristics to find the most likely keys to fit the instructional locks.

* * * * *

Identifying the Factors which Contribute to Success or Failure, in Brief


Inspect record of test performance
Use aids which link instruction and results
Study various sources of data which link instruction and results
Theory, logic, rules of thumb
Infer the nature of the relationship between instruction and results
State hypotheses


METHOD →
ULT
NAILS

CHAPTER XXI

Disciplined Creativity: Extracting Design Principles

After you have collected tryout information, you must use your creative intuition and unbridled imagination to hypothesize about the factors which made your program succeed or fail; at the same time you must also employ discipline in your thinking and ask yourself if you believe in your hypotheses to the extent that you would use them as the basis for the revision of old units and the creation of new ones.

> At some point in your analysis of constructive evaluation data you will begin to trust your hypotheses so much that you will be willing to apply them as if they were principles of design.

There are limitations to generalizations that have been made after a tryout or two, but some interesting, insightful and often valid relationships may be found. But when can you start to believe your hypotheses? The strength of your beliefs should depend upon the weight of evidence, the source of the data, the size of the samples from which the evidence was gathered, and the number of times the phenomenon has been observed. When you are convinced, use the hypotheses to guide your revision and creation.

CASE

Forming Trustworthy Hypotheses

Langbourne Rust, a consultant to the Children's Television Workshop, reported on a series of studies done on two types of productions at the Children's Television Workshop: "Sesame Street" and "The Electric Company." (1) The studies were designed to search for, define, and validate factors in program segments ("bits") to which children respond by paying varying amounts of attention. The purpose of the research was to derive reliable descriptive attributes which could be used to guide writers and producers in their programming of successful shows.

Langbourne Rust gathered data on the distractibility of five pilot segments of "The Electric Company" from a small sample (14) of second and third grade children; he wanted to know which segments attracted the children's attention and what did not. Next, he identified fifteen of the segments which attracted the most attention and fifteen which attracted the least attention. He scanned the list to find which factors or attributes were common to attractive segments, which were common to unattractive segments, and which clearly differentiated between attractive and unattractive segments: (Table)

TABLE 1

Scan List: The 30 Bits with Highest and Lowest Relative Attention Scores[a]

Name of Bit	Show	Duration[b]	Percent Attention	Standard Score
Credits	1	1	94.9%	1.79
Phone sightword	2	2	98.8	1.73
Short Circus "e on the end"	1	13	89.9	1.45
ALK Monolith	4	3	100.0	1.43
PH caveman animation	2	4	93.6	1.32
"In your own words" court scene	3	2	94.0	1.32
f, fr, ph Marquee	2	7	91.2	1.30
Short Circus "You can make up a word"	3	26	93.4	1.26
ALL monolith	4	4	98.7	1.23
Energy bridge	3	2	92.9	1.21
G sounds contest #1	3	16	92.6	1.18
2 Cosbies chip/chop	1	9	86.0	1.18
Grapefruit animation	2	6	91.9	1.18
Theater in the Dark: Gus	3	7	92.2	1.15
Movie set: "All for one..."	4	16	97.9	1.11
Credits	5	2	41.7	-3.25
Last word	5	1	43.5	-3.10
Julia Grownup	4	39	74.1	-2.58
Gag after Reasoner	1	2	30.8	-2.57
Opening song	4	11	76.2	-2.25
Cosby & Crank, f/ph	2	6	50.4	-2.07
Gag	1	1	38.5	-2.05
I am cute very, animation	5	4	58.3	-1.92
Phil on the phone, animation	2	5	52.8	-1.88
Crank call: quotation marks	5	9	61.1	-1.70
Blow/grow/throw	3	3	63.5	-1.67
Fargo North: go get gas	3	21	63.6	-1.66
Cosby & Crank: hard g/soft g	3	13	63.7	-1.65
"For" animation with DJ	2	4	56.4	-1.60
Man in the street: uncle	5	6	63.0	-1.55

[a] Relative attention scores are derived from the raw percentage attention data and express the difference of a bit's appeal from the average for the show in which it occurs. They are calculated by subtracting the percent attention to the bit from the average percent attention to the show and then dividing by the standard deviation of bits in that show.

[b] Duration figures reflect the number of 7½-second periods over which the bit extends.

Rust stated nine different hypothesized attributes. There were six attributes related to attractive segments: functionally relevant action, strong rhythm and rhyme, on-stage correcting of verbal performance, "do it one better" theme, and electronic bridges. There were three attributes related to unattractive segments: comprehensible spoken script, message monologues, and starting/ending bits.

According to his test results, Rust discovered that some factors which were not appealing: animation, music, liveliness, length of segment, and character. Rust discussed the attribute of character:

> "The identity of a character from bit to bit does not seem to affect the appeal of those bits directly. This is so even when that character has been in very unappealing bits previously. Bill Cosby, for example, participated in some of the worst bits of all, but when he was in a good role, children attended to it. While making this point about identity, it should be stressed that characters do make an immediate difference in appeal. Who they are is not important in the sense of what they have been seen to do before. But who they are is important in the sense of what they do right now. In a sense, then, children appear to be forgiving of bad roles--they won't hold it against an actor, but they are equally forgetful of good roles--it will not help a bad bit to put in a previously popular actor. The only way that would help would be if the actor changed the bit, or changed his role in it. If Easy Reader were to play Fargo North's role, the children would like it no more than they did (unless, of course, he introduced an air of more functional action). And if you could get Crank on stage to play All for one and one for all, he, too, might be a hit." (2)

Rust suspected that the characteristics of each segment were not the only factors which contributed to attraction. He discovered that segments with similar degrees of appeal followed each other 2.6 times as often as did segments with different degrees of appeal. There was a consistent relationship. Thus, each segment was influenced by the one which preceded it. But the influence extended no further than one segment, and no further than one minute.

Rust pulled these ideas together into a set of hypothetical statements to be used to predict (and possibly influence) appeal:

"1. If the bit lasts one minute or longer, compare the numbers of high- and low-appeal intrinsic attributes it possesses.
2. If the bit lasts less than one minute, take its own intrinsic attributes together with the intrinsic attributes of the preceding bit, and compare the total numbers of high- and low-appeal attributes.
3. If there are more high-appeal than low-appeal attributes, estimate a high level of response.
4. If there are more low-appeal than high-appeal attributes, estimate a low level of response.
5. If there are equal numbers of high- and low-appeal attributes, or if there are no intrinsic attributes at all, make no prediction." (3)

These are the slightly modified definitions of attributes which Rust referred to in his hypothetical statements:

> "Functional action. Bits that portray locomotion or active movement through space that is directly functional to the development of the plot or theme of the segment. Pointing, writing or arranging things by hand do not qualify; neither do movements that are not directly functional to the plot (such as walking around in order to switch scenes).
>
> The bulk of the segment must portray this functional action, be in very obvious expectation of it, or in clear reaction to it.
>
> Strong rhythm or rhyme. Bits in which strong repetitive rhythm and rhyme occur together, for most or all of the segment in question. These qualities may be present in songs, verse, or 'jive' talk.
>
> Portraying children. Bits that involve children, or animated child characters, on screen for most or all of the segment.
>
> On-screen disagreement. Bits which have a theme of one character's attempting to correct another on reading, pronunciation, or writing. Both characters must be on screen.
>
> Repeated attempts. Bits in which the central theme is one of repeated attempts to achieve some concrete goal or standard. The standard may be set by a competitor's performance, by the performer's own achievements, or by some other concrete criterion which is made clear to the audience.
>
> Comprehensible spoken script. Bits that have a spoken soundtrack that is comprehensible without reference to the screen. The whole meaning of the

> bit need not be auditory, but the auditory must make sense on its own. Telephone conversations usually have this attribute. This definition does not include bits involving the slow sounding-out of letters (blending).
>
> Message monologues. Bits in which there is only one character throughout, and where that character is on-screen in a more-or-less stationary position, telling the audience something (reading to himself does not qualify). This definition does not include bits where the message is directed at other characters.
>
> Program identification. Bits that are devoted to the identity of the show or information about it: show number, name, theme, credits, etc." (4)

With this set of rules, Rust was able to account for the results in the data collected for shows 1 - 5 of ninety-four bits correctly and eighteen incorrectly. This yields a prediction ratio of 5.2 accurate predictions to 1 error. Using the rules with some minor changes, he was able to predict the appeal of shows 6 - 10 to six children, with a prediction ratio of 4.50 to one.

To find out if other reviewers could predict using the defined attributes, Rust showed them videotapes of "The Electric Company" and asked a pair of reviewers to rate the segments after only reading the definitions of the attributes. The pairs of reviewers agreed with each other, on the average, eighty-seven percent of the time. With seven attributes (one was found not to contribute much) the reviewers predicted correctly one hundred and seventeen times, incorrectly, thirty-nine times, and refrained from prediction one hundred and sixteen times. This level of prediction could only come about by chance once in a thousand tries.

These results were compared to results obtained by reviewers who made predictions only based on their own familiarity with the distractor measure. The experienced reviewers made one hundred and fifty-two correct predictions and ninety-three incorrect predictions. The level of prediction could only come about by chance once in a hundred times. Thus, predictions made by individuals on the basis of identifiable attributes may be as good or better than experienced reviewers.

At this stage, Rust felt convinced that his hypotheses were capable of being used as design principles. Rust concluded,

> "The most direct implications, perhaps, are that writers and producers of 'The Electric Company' should strive to embody the high-appeal attributes and seek to avoid the low-appeal attributes in their new programs. The guidelines they provide are not exhaustive: material that embodies none of the discovered attributes may be highly appealing to children (or very unappealing); but where they apply, they should be heeded. If visual attention to the television screen is desired, one should avoid the low-appeal attributes. If one wants to be certain of high attention, building in the high-appeal attributes will help." (5)

Summary

Do you trust your tentative conclusions about the strengths and weaknesses of a tested unit so much that you are now willing to risk using them as the basis for making some decisions? That is the question.

CHAPTER XXII

Metamorphosis: Generating Modifications

Improvements of instructional programs do not happen by accident nor are they natural occurrences like the transformation of a caterpillar into a butterfly. A project director must take into account the inferences drawn from the data collected and he must proceed systematically to remove program faults and develop program strengths.

First, decide if revision is necessary.

Although some evaluators suggest that you revise whenever an objective is not attained, (1) (2) it is not always that simple. For example, if your analysis shows that a student is failing because of inadequate use of materials, as opposed to inadequate materials, you should not have to revise materials. Your decision to modify a unit depends on the importance of desired results and the degree to which those results have been achieved. To assist you in evaluation, you should class your results as substandard, good, and excellent; a substandard result, for example, might be that half the students have not achieved the objective.

If very important objectives are classed as substandard, you should certainly revise portions of the program related to them. You should not revise sections associated with objectives for which recorded results are good or excellent. But you may revise unimportant objectives showing substandard results or very important priority objectives showing only moderately good results.

You revise programs with unimportant objectives showing substandard results if the program is producing counter-productive side effects--it makes the students, for example, hate math. Revising such a program also depends in part on the extent to which you can estimate the possible gain to be made by a revision.

When you estimate gain, look for the chances of obtaining more of the desired behavior, a closer approximation to the desired behavior, less undesirable behavior, fewer counter-productive behaviors, less irrelevant behavior, more efficient behavior, or greater enjoyment of learning at lower cost.

Consider the quality of your data, and the resources necessary to achieve part of that gain. In addition to gain in terms of achievement you can ask, "Can the program be produced less expensively, be packaged better, or made more consistent, or easier to use?"

There seems to be a point of diminishing return when trying to reach a certain criterion. The first test revision cycle may cost you a certain amount and result in a jump from fifty percent attainment to seventy percent for most students. But it may cost many times as much to improve from seventy percent to ninety percent. (3) You must decide how much additional effect is worth the additional money for revision.

Revisions are also needed because of constraints. Too little money may cut a program or the lack of available talent may force the change of a sequence of instructional films. There may be budgetary, logistical, and technical factors forcing modification, and many of these needed changes could come during a review of early plans. You may, for example, need revision to add to the polish of your materials. You should decide to revise while you still have time. The revisions may be used for other similar units.

If you do not take into account the factors listed above in considering the necessity for revision, you may find yourself in a frustrating situation. You may decide, for example, to revise because of one factor--students have trouble on post-test items and some report a negative attitude about the topic. But you should consider other factors because you may receive considerable criticism for spending time and money on what may seem to be an unimportant objective. You may also find that you do not have enough money to produce the revision well, and you do not have enough time to create the next version for a tryout. You may also find that you have relatively little to gain considering the cost.

CASE

Deciding if Revision is Necessary

> Remember the psychology professor who found that his students were not learning the principle of negative reinforcement from his study units in psychology? After collecting data and hypothesizing about the reasons for the results, the instructor thought that revision was necessary. He cited five reasons: 1) the results were well below the standard set and 2) because learning the principle was important. 3) There was also considerable gain possible from sixty percent (the present level of achievement) to a possible ninety percent, and 4) the units were not rough enough for the instructor to believe that the lack of polish was responsible for the lack of achievement. 5) Finally, the most likely revisions would be additions and they should not be too costly and might be completed quickly. All of these ideas convinced the instructor that revision was needed.

You may hold a segment for later consideration, cut it, add new parts, add more portions found to be successful, or change its quality.

Your analyses and hypotheses determine your revisions. Add more of something when you find not enough is available. You need a new approach when evidence indicates that the approach is inadequate. You eliminate components when they are found to be irrelevant or interfering. You create a qualitative change when students are misled. You maintain and duplicate a component when evidence shows it is making a positive contribution.

You may decide to rest a faulty segment and try again at some later time, or you may decide to eliminate a faulty segment. For example, when researchers find a "Sesame Street" segment frightening to children, they may eliminate it. If they find a segment scoring very low on attention measures, they may withdraw it from one show and try it on another later. But elimination of a segment, simply because it is hard to measure, is a form of retreat. (4) The decision to drop a segment or an approach must be based on strong evidence and logical argument.

With some media, notably film, you may wish to avoid drastic cuts at first because it is harder and more costly to re-edit film. (5) Consider presenting your message in various ways--in a magazine format, for example. The great advantage is that when you are forced to eliminate one piece, not all is lost.

You may decide to add to a method or product by use of new, better designed components or by adding proven components. For example, if animations are found to contribute to learning, and more learning is required, you may add more of the same cartoons, or you may make more cartoons of the same type. Changing the components of a unit qualitatively requires the most work. If a segment is found to be ineffective and is still necessary, you may have to redo it.

Students can suggest modifications.

Producers who want to limit their use of theory in making revision could depend on student suggestions. Students can provide good ideas for revision.

CASE

Collecting Student Suggestions for Revision

Abedor collected the following comments made by students after a lesson on cattle breeds; you will find many suggestions for revisions and some statements describing problems that could easily be turned into revisions. (6)

"1. Too much new information too fast.
2. Slides don't exemplify the specific breed being talked about on the tape.
3. Poor example of specific breeds; e.g., the 'Red Poll' was brown and the 'Black Angus' was navy blue, a horned breed was shown without horns.
4. Should use simultaneous, not sequential, presentation of different breeds.
5. Overemphasis on historical development.
6. Critical cues not highlighted on pictures of different breeds.
7. Use more than one shot or example of various breeds.
8. Graph in workbook totally unfamiliar and unusable.
9. Workbook has insufficient space to take notes.

10. If a slide is omitted because there is not a good photo of a breed--tell the students.
11. Have students write own definitions in workbook.
12. Make alternate forms of the pre- and post-test.
13. Do not use black and white pictures of colored breeds.
14. Break the lesson into two parts, foreign and domestic breeds.
15. Exams don't reflect lesson content."

CASE

Collecting Student Suggestions for Revisions

The Far West Laboratory reported revisions in their mini-course, Discussing Controversial Issues, which were based on student suggestions.

> "In response to student comments on the course, the Student Handbook has been rewritten to incorporate cartoons in an attempt to make the reading more interesting. In addition, the reading level was lowered and humor was added. The writing style became more direct and informal. Students indicated that they disliked the model tape check list discriminations. Accordingly students are now asked to watch for certain discussion characteristics, and the model tape is intended to stimulate discussion." (7)

Teachers can suggest modifications.

To find appropriate revisions, attend to and use teacher requests. (8) Their comments are usually valid, and by using teacher comments an evaluator may be able to gain teacher rapport.

CASE 1

Collecting Teacher Suggestions for Revision

Roger Scott, a product developer at the Southwest Regional Laboratory, reports about teacher comments which influenced revision in an early tryout of the Instructional Concepts Program.

"The changes made in the program originated in the reports of participating teachers, classroom observations by SWRL staff, analyses of student test data, and analyses of teacher questionnaire data."

"...The most important change in the instructional program involved the materials in Unit 1. The first three lesson taught a total of ten color names to children and each lesson included a teacher-read poem. Teachers reported that the poems were difficult to read and were confusing to the children. Many teachers also expressed the desire to begin the instruction with a slower learning pace. Accordingly, there are only two colors per lesson in the revised material and each lesson includes a story rather than a poem.

The Program Resource Kit containing all of the stories, flashcards, games, daily assessment cards and criterion exercise directions, was completely reorganized at the request of teachers. In the tryout all of the games and flashcards for a particular unit were sequenced together. In the new version, a few game cards and flashcards are placed directly behind the story card and daily assessment card for each lesson." (9)

In a later test of the concept program Scott reports:

"Most of the concepts taught in the original Instructional Concepts Program are included in the revised program. Four concepts relating to prereading skills were added, since the program is used before children receive any reading instruction. The concept "not" was added at the request of teachers and curriculum specialists in the tryout schools. Although teachers liked the lessons dealing with pattern they agreed that these concepts were not critical for future academic performance and consequently they were dropped from the objectives."

"The revised program is divided into seven units. Unlike the original version, each unit contains concepts related to a single dimension such as color, size or amount. This was done at the request of tryout teachers who felt that such an arrangement would facilitate evaluation of student performance and scheduling of additional practice. The units were sequenced according to pre-test data. Scores were highest on colors, so that unit was scheduled first; the next highest scores were on sizes, so the unit on sizes was sequenced second, and so on." (10)

> "Practice exercises were also found in need of major revisions. In the original program a single page which illustrated the concepts to be taught was included for each lesson. This was an optional activity which teachers could hand out and ask children to color or mark. A number of teachers suggested that this component was not structured enough to be useful in the class. Because of these comments and because of a desire to coordinate the program with the SWRL Communication Skills Materials, the practice exercises were completely revised. Each revised exercise consists of four pages with each page divided into five rows. Directions are printed in the margin of each row so that they can be read from the left hand side of the paper. These directions, which can be used by the teacher, an aide, a parent, or a tutor, ask the child to identify illustrated concepts by pointing and naming." (11)

CASE 2

Collecting Teacher Suggestions for Revision

Morris Lai, a product developer at the Far West Regional Laboratory, took into account teachers' comments in the revision of a unit called "Discussing Controversial Issues":

> "Because teachers complained about the rigidity of the four-week schedule, the revised course was made self-pacing. Each teacher will decide how long to spend on a lesson.
>
> Sample lesson plans were developed, based on what field test teachers said seemed to have worked the best. They provide guidelines for planning activities with students and suggestions for using the course materials, choosing topics, giving feedback, and giving assignments that maintain students' interest." (12)

You can use intuition, insight, and a good dose of common sense to generate modifications; but behind most decisions is a set of empirically or theoretically based ideas.

On occasion, your hypothetical reasons for a program's success or failure will be easily converted into revision: if a fault seems due to too few examples, then you add examples; if a fault seems due to lack of practice, add practice exercises. But when the translation is not apparent, theory plays a key role in modification. Consider what you want to have happen, what you have to possess to make it happen, and then use the principles which relate contributing factors and results.

CASE

Considering Theoretical Principles for Modification

Here are some examples of theoretical principles from varied sources.

Here is one from advertising research:

> "...It is quite well established that meaningful material is better remembered than meaningless material. The brand cue must trigger an interconnected structure of recollections. The more meaningful the structure is, the better chance it has of surviving a night's sleep. There are many commercials created to catch attention. However, those very attention-getting devices are often absurd from the standpoint of the viewer--absurd in the sense of having no meaning in relation to what is being advertised. Therefore, it is not surprising that when we call the next day, she cannot remember seeing a commercial for that brand. It is only doing half the job to get people to pay attention. You must also communicate with them in a way that meaningful to them." (13)

Here are some from "Sesame Street" research:

> "Beyond these useful diversities in characters, content, and style, varied pace and mood are critical in sustaining attention. The appeal of any single segment is tied closely to the contrasts provided by the episodes preceding and following it. Both fast-paced and slow-paced material will hold children's attention (the common criticism that Sesame Street is continuously frenetic simply is inaccurate), but a slow, peaceful episode is more appealing when surrounded by fast-moving episodes

> than when it follows another slow, quiet piece. Interest in any particular episode is higher if it creates a pace and mood that looks, sounds, and feels different from the one that preceded it. The principle that visual action and contrasts appeal to young children need not mean that the action must always be rapid or frenetic to be effective; instead, the pace of the action should be varied." (14)

Here are some from "Electric Company" research:

Research results suggest rules for producing electronic bridges on "The Electric Company." Electronic bridges rearrange the same set of letters to form different words (bat to tab, tool to loot, chin to inch). The basic well-documented design principle behind the suggestions is that varying the minimum number of sounds and symbols will teach a child to recognize the difference between two words.

> "1. Do not separate consonant diagraphs or vowel combinations; they are being taught as a unit.
> i.e. shore/horse, plate/pleat, eat/ate, seam/same, sheet/these, are not acceptable.
> 2. Do not have a letter silent in one word and pronounced in the other.
> i.e. are/ear, lame/meal, plane/panel, evil/live: be consistent.
> 3. Make sure the sound of the letters is similar in both words.
> i.e. ocean/canoe, raced/cedar, would be too confusing.
> 4. Avoid exceptions.
> i.e. stake/steak" (15)

Regarding the appeal of "The Electric Company" as a function of the time within the show, the show's researchers state:

> "The point is that children do not automatically pay attention to the last part of the show. It has approximately the same average score as the middle section. However, relatively few of the less popular bits and most of the very well liked ones such as Letterman and Very Short Books tend to appear in the last third and to pull up the attention level there."

> "This study yields a number of implications. First, in order to raise the appeal level of the whole show, it might be advisable to intersperse bits

> of known high appeal throughout the show, not put them at the end.
>
> Second, the low scores at the beginning of the shows should be read remembering that the distractor slides are at their most diverting at the beginning of a show. Scores there are usually low, but some of this lack of appeal is due to the novelty of the slides." (16)

CASE

Using Theoretical Principles in Revision

One question that producers of "The Electric Company" asked after looking at data and talking to researchers was whether or not slowing the pace of a show would make the show more comprehensible but less appealing. Researchers at "The Electric Company" changed the speed with which words were said and shown during the show to produce a "slow" show. The results in terms of attention, comprehensibility, and achievement were studied and, in cases where results for "normal" shows were available, they were compared.

The show was about average in the amount of attention recorded. The distractor percentage for the whole show was 76.7%. (Children watching the show faced the screen most of the time.) More than ninety percent of the children who watched the show were able to answer correctly all but two of the questions about events in the program. Here are examples of specific reports on responses to questions.

> "They knew what the amusement park was and could enumerate some of the things they had seen. Most of them even thought they recognized the location (Coney Island, Palisades, or Roseland). The merry-go-round and the cotton candy were most often remembered."
>
> "All of them understood the "Sit" sequence with Paul. They knew what the sign said and they realized that Paul kept guessing the wrong word until he finally read the sign correctly." (17)
>
> "All but one of the children knew the word at the end of the snap bridge."
>
> "All except one child could tell the story of the Bee on the Knee animation." (18)

A normal show could be as comprehensible as the slow show, but could not be much more comprehensible.

> "Of the eleven children tested on all 18 words in the show, only 3 knew a word before the show that they no longer knew after the show; i.e., there is a good chance they were guessing on the pre-test."
>
> "Of the remaining 8 children, 5 knew only a third of the words to start out with, and the other 3 knew even fewer (3-4 words). None of them were guessing. Every one of the children learned at least one word during the show. Two of them learned 2 words each; two others learned 4 words each, and one child even learned 6 new words." (19)

A "slow" show can gain the attention of children, at least as well as a "normal" show and a "slow" show is as comprehensible as a normal show could be. But one might feel more certain if averages of comprehension measures for normal shows were used for comparison. This data is being gathered. One might believe that students do learn from a "slow" show; that seems credible when you consider that eleven children, as different as children can be from each other, who could not read words when asked, were able to read them after a half hour experience. But, we might feel more sure of the results if the show was compared to a placebo experience and a normal show. This data is being gathered, also.

You can use revision tryouts to confirm principles.

CASE

Confirming Principles in a Revision Tryout

Silberman and Coulson used a revision tryout to confirm principles. (20) They developed instructional programs by use of tutorial procedures. After four programs were developed in this manner, the producer hypothesized that three principles were responsible for faults found and were the basis for remedies in all programs. These principles were gap, irrelevancy, and mastery. Gap meant that specific information for each criterion item had to be included. Irrelevancy meant that information unrelated to criterion questions should be cut. Mastery meant that students were required to demonstrate learning on one subject before proceeding to the next. To verify these principles, programmed texts were developed with and without the principles, and, when principles were not represented in the texts, performance suffered.

Silberman and Coulson created six variations of a logic program. The complete logic program used the three principles, gap, irrelevancy, and mastery. The other programs contained combinations of gaps, irrelevancies, and left out the branching contingencies required for mastery. The first variation was the good version, containing student diagnostic tests which required responses; based on his responses a student is given remedial work and another test. The second version was the linear version containing no branching. The third, the small-gap version, was like the linear version with some items either changed or deleted. For example, one of the two items deleted was that the truth or falsity of premises and conclusion of an argument do not affect its validity. The irrelevant version, the fourth variation, was like the linear version, but two irrelevant items were added. For example, students were told about truth tables and Latin names for forms, material not required on the post-test. The bad small gap version, the fifth version, combined the linear, small gap, and irrelevant versions. The sixth variation was the bad large gap version like the bad small gap version except that another gap was included.

Ninety-one students composed the six groups taking the tests. They all took a post-test consisting of material consistent through all six programs and of material modified in different programs.

In this case systematic elimination of factors from a program confirmed some of the ideas hypothesized by developers after doing constructive evaluation.

Silberman and Coulson concluded:

> "In short, two of the three independent variables, gaps and irrelevancies, had a significant, cumulative and specific decremental effect on post-test performance. These effects were not obtained at the cost of giving the good groups added training time; if anything, the data suggest that the groups who took the greatest amount of training time received the lowest scores on the portion of the criterion test covering the program segments that had been experimentally modified."
>
> "While it is possible that the addition of remedial branching does not improve a linear program, as a comparison of the good version and linear version scores would indicate, two alternative explanations are possible. First, it may be that the diagnostic questions used to determine the need for branching did not assess the difficulties students encountered on the post-test. Second, it may be that the remedial items used were not adequate to overcome the students' lack of learning." (21)

Usually the use of theory to create revision takes place in the head of an instructional developer. It is a rare event to find that someone has written down what principles he has applied.

CASE

Stating the Principles Used in Revision

Roger Scott of the Southwest Regional Laboratory wrote down what principles he applied to create revisions for the preschool concept program described in earlier chapters. The example also includes use of teacher suggestions.

> "Early in the tryout, it was determined that the format of the story illustrations would have to be changed. One poster was used to illustrate each of the stories in the revised program. Three cards were used to illustrate each story in the original program, but teachers reported that the posters were cumbersome. Lesson observations by SWRL staff also indicated that the posters were used in a manner which prevented children from frequently practicing the use of the concepts in the lesson. Teachers typically asked individual children to come to the front of the room and point to an instance of the concept illustrated on the poster. With a lesson conducted in this manner, many children did not have a chance to engage in appropriate practice. [Practicing the precise task specified in an instructional goal is an important theoretical instructional principles.] Others had only a very limited opportunity. In order to increase the frequency of practice, concept books were developed for the revised program. All children received a book for each of the program's seven units. These books are similar in format to the storybooks used in the SWRL Communication Skills Program. Each lesson is illustrated on two pages which face each other. The illustrations include the unit theme character and objectives familiar to inner city kindergarten children. The illustrations also represent two or more instances of each concept included in that lesson. Concept naming and identifying questions to ask the class are listed in each book." (22)

CASE

Using Theory in Revision

In the case of the psychology teacher whose students had not learned the principle of negative reinforcement, he used theory and student suggestions to generate his revisions. His hypotheses were that there were insufficient explanation and demonstration, lack of differentiation between punishment and negative reinforcement, insufficient examples and practice, inappropriate practice, and lack of agreement on the definition. Reversing the hypotheses into solutions is easy but not complete without application of theory into practical procedures. The psychology teacher should provide more explanation and demonstration, differentiate between punishment and negative reinforcement, present more examples, allow more appropriate practice, and resolve the disagreement on the definition.

But how should all this be done? The teacher decided to do most revision in the form of handouts and to introduce some changes in his lecture. The handout changes were primarily additions, some of the same things already used and some new ideas. The lecture required qualitative changes.

The first handout included an explicit definition of the principle and an explanation of the reason for the difference of definition in the text. The various contributing factors and the dependent variables in the principles were each stated, diagrammed, and compared to the principles of punishment and positive reinforcement. Contrasting examples of each were present. References were made to common bits of knowledge which illustrate the principle like "The Taming of the Shrew" and the story of Solomon and the two mothers.

The second handout included practice in discriminating between negative reinforcement and other principles. Examples of each were given, and students were asked to label them just as they would in the test.

A third handout included cases in which either the principle of punishment or negative reinforcement is suitable. The student must decide which is correct for a given case. Also included in the third handout are cases for which students had to write prescriptions applying principles, many of which required negative reinforcement. This practice was the same behavior required for the test.

The lecture plans followed the handouts. Students were told to read the first handout before the lecture. During the lecture, the teacher was to present several cases and demonstrate how he would apply the principle of negative reinforcement. Then, within only ten more minutes than he usually devoted, he was to give students class practice in solving similar cases and let them check each others work.

Creating a revision is solving a problem. Therefore, a producer could benefit by applying procedures which are used to make problem solving easier.

You should follow some problem-solving strategy. You might attack one segment at a time, produce a detailed definition of the problem, and search for several solutions or partial solutions for the same problem. You might first handle revisions for all major problems (those indicating changes to objectives, sequence, content and tests), and then work on minor ones (examples and better instructions). You might use these problem-solving heuristics to generate revisions:

1. Think about elements of a problem several times.
2. Vary the relationships of the elements by creating a model or a drawing.
3. Produce more than one solution before you act.
4. Talk over the problem with someone.
5. Use group resources; ask for other views.
6. Evaluate your ideas carefully before you act.
7. Delay choice of a solution until you must act.
8. Stop when you are stumped and come back to the problem later.

Most of the heuristics are designed to avoid jumping to conclusions.

CASE

Using the Heuristic of Delaying a Decision

At the beginning of "Sesame Street" children were not learning much from the game "One of these things is not like the others." Had the producer eliminated the segment he would have made a mistake. Children simply needed time

> to learn the way the segment teaches; then they began to learn the content. Producers observed a similar phenomenon with the detective, Fargo North, Decoder, on "The Electric Company." Once children could understand his word decoding routine they began to learn from the segments. (23)

Producers may feel that theoretical principles will reduce their creative options, but it is more likely that principles will create new frontiers and that lack of principles may stifle creativity and set limits to a producer's creativity. Langbourne Rust commented on the limits imposed on producers when principles are not available:

> "One effect of being able to delineate attention-controlling attributes is to permit television production to be much less conservative than it has been in the past. Not knowing just what it is about successful shows that makes them succeed, television producers have tended to work within very narrow limits, creating 'new' shows as similar as possible in every conceivable way to a demonstrated winner, varying only far enough to establish an identity separate from the model's." (24)

Principles provide the basis for a creative act. There are principles and elements of visual design, and I find that their existence does not disturb most visual artists. They use the design elements as foundations and as a set of evaluative guidelines, and it seems visual artists have not yet run out of creative possibilities.

Usually not all changes can be put into effect because of the limits of existing resources; you must make priorities and select among modifications to be put into effect.

You may produce more ideas for revisions than you can use. You must then determine the order of priorities among your list of modifications. Some individuals--a team or a producer--may have the final say on which changes are made. Those individuals who make the decisions must have the authority to spend time and money within limits because making revisions means spending additional money and wasting money that has been spent. That is why many revisions are impossible for small scale projects.

To determine the order of priorities among a list of modifications, each suggestion is compared to the following criteria, and decisions are made:

Priorities are given to revisions

1) of lower cost. For example, the psychology instructor wanted to incorporate most of the revisions into lecture, but he had little extra time. He did have some funds for printed materials, so he settled for that.

2) with a minimum effect on other unrevised parts of the program. For example, some programs have many elements, text, practice workbooks, visuals, and tapes. Change any one of them, and you may have to change all of them. In many cases you may have to modify the whole program or leave it alone.

3) within your production capabilities.

4) which are low in cost and take little time to complete. The greater the cost in time and money and the tighter the time schedule, the more likely minor faults are to be left in. When production is behind schedule, changes are less likely.

5) which are data based. Someone must keep a cool head, remember to make revisions based only on what the data showed needed revision, and check to see that all needed revisions are made. Otherwise, a good many revisions can fall by the roadside.

6) which give the most effect for the cost. For example, by a few handouts the psychology teacher could make a great change in learning. One way to determine if the charge will be worth the cost, is to check to see how many students reported the problem as an important one to be remedied and how many sources of information indicate the extent and influence of the problem.

7) suggested at a time when the material is most changeable.

8) of media which are easy to change. For example, changing pencil and paper is easy; changing videotapes or film is difficult.

9) acceptable to producers, administrators, and reviewers.

10) which leads to achievement of important objectives. A good revision helps students to reach the program goals better than the previous draft. To improve is not just to remedy faults; it is also to expand on the positive possibilities of the program.

11) which are theoretically sound.

Summary

After you decide a revision is necessary, collect suggestions from students, teachers, and your staff. Balance the amount of intuition, problem solving, and the amount of well documented principles and theories (confirmed by previous research or by a revision tryout). Then when you must finally decide to put revisions into effect, make priorities and select among the revisions to be made.

Generating Modifications, in Brief

Decide if revisions are necessary.

Consider

- student suggestions,
- teacher suggestions,
- intuitive impressions,
- theoretical application (which you can confirm in a revision tryout),
- problem solving procedures and heuristics,

when you

- hold a segment for later consideration,
- cut it,
- add new parts,
- add more portions found to be successful,
- change its quality.

Make priorities among revisions to be put into effect.

CHAPTER XXIII

Try, Try Again: Recycling

Testing revisions can provide useful information about the quality of changes made in a program and about the need for further improvement.

Recall the old saying: "If at first you don't succeed, try, try again." If you find that your materials do not succeed at first, you should revise, and then test the materials again. This procedure is commonly called recycling because you proceed again through the entire constructive evaluation cycle.

By recycling you can check the effectiveness of your revisions and explore the need for further improvements. But few instructional developers do retest; they simply assume their revisions work. The reason that the evaluation process is not often repeated after one revision is that the producer is tired or that the evaluator is unable to repeat the evaluation for lack of time or money. It is interesting to note that the reason for not evaluating again is not that the changes the producer made resulted in greater achievement; usually there is no data collected to substantiate such a claim.

After you make revisions, you must decide if you should retest the new version of the instructional units.

To see if a retest is appropriate, you may consider these factors:

1. The time remaining until the method must be used to teach.
2. The money remaining.
3. The freedom given to producers to revise.
4. The effort required for retesting and making additional revision.
5. The nature of the modification made.
6. The achievement results recorded during the first tryout.
7. The doubt left in your mind.
8. The importance of the goals.
9. The other jobs which must be done.
10. The pressure imposed by administrators or sponsors.
11. The access to new information. (1) (2) (3)
12. The need for evidence to convince people that the program works.

You decide not to retest if you have little time, money, freedom to revise, or access to new information. You do not proceed if you cannot scrap what you have produced or if another tryout requires more effort than the first. You do not retest if you recorded achievement results on the first tryout close to criterion, or if the goals of the unit were relatively unimportant. You do not retest if you have no doubt about the effectiveness of the program, if other jobs are pressing, or if administrators are demanding completion. You do not retest if you are required to add segments, similar to tested successful ones, a unit separately tested, or eliminate portions.

When there is time, money, and freedom to revise a program, you test again. You can go ahead with another tryout when you can still scrap segments, when many extensive modifications are being made, or when recorded achievement results relating important goals are far from criterion. If the effort to produce a second tryout is about the same or less than your first one, if there are few other jobs to do, or if administrators are not demanding completion, you can retest. You can conduct a retest if there is doubt about the success of the program left in your mind, if new information may be forthcoming, or if you are making qualitative changes or adding new segments.

The decision not to go ahead with another constructive evaluation tryout does not mean that you cannot collect more data. You may be interested in collecting more information for reasons other than improvement. You may wish to convince others to use the program or convince sponsors to provide more money.

Usually a first tryout includes one unit, but if you decide to retest the new version of a unit the second tryout may include 1) the revised unit only, 2) the revised unit and similar but untested units, or 3) for comparison, the revised unit and the original version.

In a comparison of revised and original materials, you should be cautious about favoring the revised materials. For example, if objectives or criterion test items change from the original to the revised version, it is not fair to use a test made only for the revised program.

A second test will require changes in the tryout elements.

When you recycle you must choose new tests, samples, instructional units, and test sites. You should choose samples consisting of groups rather than samples of individuals; you should choose large groups instead of small groups. You should question new people if your choice of the original test sample was inappropriate, if your results suggested that the program would teach other audiences, or if you want to be sure of the validity of your results relating to a particular audience.

You should test a first draft in a laboratory site, but you should test a second draft in a field test site. To be fair in a comparison test, you may think you should use standardized tests. (4) When using a standardized test to compare programs, you are likely to find no difference between the results even if real differences exist, because it is likely to be an insensitive, unrepresentative, low fidelity measure.

You can select or create a specific test for each program and then combine the two tests into one which will possess items common to each program and items unique to each program. (5) A combined test provides the advantage of comparing the merits of two programs on common objectives and also finding their individual contributions.

To make a comparison worthwhile you have to be sure one of the drafts or programs is truly more effective than the other. (6) It is worth neither the time nor the money to compare versions with only a slight possibility for a change in results.

There is no magic number of revision-retest cycles.

You stop testing when the instruction is effective or useful enough for a certain number of students. A rule should be established that revision and testing will stop at a certain time, at a certain level of competence, or at a certain stage in production.

You should ask analytical questions when you test a revised unit.

First you ask if the results of the revised version meet your desired standard. Second, you ask if your program has improved. See if the results are nearer to the standard than they were after the original draft.

CASE 1

Asking Analytical Questions on a Retest

The results of the revised version of the math program reported by Judy Light showed that students reached criterion on twenty-seven out of fifty-five objectives, improved on eighteen, remained the same on seven, and did worse on three. She states that student performance improved on eighty-two percent of the objectives analyzed. (7)

CASE 2

Asking Analytical Questions on a Retest

Abedor compared original and revised versions of instructional units in cattle breeding and reported that post-test achievement scores were 1) not at a satisfactory level, 2) showed marked improvement, and 3) were significantly better statistically than the original. (8) A large percentage of the students achieved the eighty percent criterion required by the instructor. In some units one hundred percent of the students achieved the set criterion. Gain scores from pre- to post-tests were better in two out of three units. In some units students reached the criterion in forty-seven minutes on the revised version as compared to 42.85 percent of the students reaching criterion after one and a half hours on the original. In one exceptional case there was an improvement of only 8.27% of the students reaching the eighty percent criterion. This, however, might have been due to test problems or due to incorrect practice cues: identification of animals was not practiced the same way in the book as on the test and the colors portrayed on the test were not true.

CASE 3

Asking Analytical Questions on a Retest

A filmstrip, "The Sun and Its Planets," was tested twice using large groups of children. (9) For every idea in the filmstrip a multiple choice test item was given. After a first draft tryout, Vandemeer, an instructional researcher, used test data to analyze the program. He related low scoring items to the filmstrip presentation and made revisions to add more cues, provide higher visibility to certain characteristics, and simplify language. Then Vandemeer tested the revised versions. He found that some of the revisions worked and some did not.

The second revision was compared to the original filmstrip. Generally, the results showed that those students seeing the revised filmstrip had higher scores on the average. They were children in grades 5, 6, 7, and 10, randomly assigned to see either the revision or original filmstrip. Thirty-five of the sixty original frames in the filmstrip were different in the revised version, twenty-one proved favorable.

The following results show that criterion was reached on the test of one original segment and no improvement was seen on the retest results.

> "The third test item required the student to identify phases that correctly describe the character of the sun.
>
> 3. The sun is a huge globe of
>
>   1. solid coal that will burn forever.
>   2. earth covered with hot lava.
>   3. rock polished like nickel.
>   4. glowing hot gases.
>
> The revision aimed to convey the impression of great heat by making the sun appear brighter and by making the margins of the sun less clean cut. Also, focus was given to the relevant information by reducing the number of irrelevant statements from two to one.
>
> Table 3 shows that almost all students were aware after seeing either version, of the characteristics of the sun." (10)

The choices refer to test alternatives. FSO stands for filmstrip original; FSR stands for filmstrip revised. N stands for the number of students.

TABLE

Table 3

Percent Choosing Various Responses to Item 3

Choice of Answer	Grade 5-6		Grade 7		Grade 10	
	FSO	FSRR	FSO	FSRR	FSO	FSRR
N	72	68	72	71	59	61
1	3	1	0	0	2	2
2	3	0	0	0	0	0
3	1	0	1	1	0	0
4	93	99	99	99	98	98

The results for test item 4 show that considerable improvement was made and criterion was reached. Test item 4 was

"How many earths side by side would it take to equal the diameter of the sun?

1. 50
2. 108
3. 866,000
4. 1,300,000"

Vandemeer's comment on the results was:

> "Item 4 calls for the student to select the correct ratio of the earth's diameter relative to that of the sun. The correct response could be made to item 4 by reference solely to the verbal elements of the filmstrip. The differences in this verbal element are 1) the heading of the revised frame alerts the learner to the huge size of the sun, 2) the actual diameters of the earth and sun are shown in the revision, and 3) the revision omits reference to the relative volumes of the earth and sun.
>
> Significant differences in favor of the revised filmstrip were found at all grade levels tested in terms of the proportions selecting the correct response." (11)

Table 4

Percent Choosing Various Responses to Item 4

Choice of Answer	Grade 5-6		Grade 7		Grade 10	
	FSO	FSRR	FSO	FSRR	FSO	FSRR
N	72	68	72	71	59	61
1	3	1	0	4	0	0
2	39	71**	58	89**	72	88*
3	26	24	21	6	14	10
4	32	4	21	1	14	2

*Significant at .05 level
**Significant at .01 level

The following results demonstrate an improvement without reaching criterion.

> "Test item 29 gets at the motion of the planets in somewhat more concrete terms, in that it sets up a hypothetical situation and requires the student to identify the appropriate response to the situation by applying information presented in the filmstrip.
>
> 29. At 9 p.m. on March 1, you see the planet Jupiter as you face straight south. If you look again on April 1, at the same time, where will you see Jupiter?
>
>   1. at exactly the same place where you saw it before
>   2. closer to the western horizon than where you first saw it
>   3. to the left of where you first saw it.
>   4. to the right of where you first saw it." (12)

"In contrast to the results found from Item 28, responses to Item 29 showed consistent and statistically significant differences in favor of the groups who saw the revised filmstrip." (13)

Table 29

Percent Choosing Various Responses to Item 29

Choice of Answer	Grade 5-6		Grade 7		Grade 10	
	FSO	FSRR	FSO	FSRR	FSO	FSRR
N	72	68	72	71	59	61
1	22	9	17	4	9	7
2	35	25	28	33	22	16
3	21	43**	28	46**	38	64**
4	22	23	27	17	31	13*

*Significant at the .05 level
**Significant at the .01 level

The analysis of the following items shows that criterion was not achieved or just barely achieved, and improvement was not evident.

"28. How can you tell the difference between a planet and a star?

1. stars are in the same relative position every night
2. stars have a slightly different color
3. stars become brighter during a full moon
4. stars are brighter than planets" (14)

> "Table 28 shows that there were no significant differences among groups of students who saw the alternative versions in terms of their responses to Item 28. Only in the case of the tenth graders did the majority of students respond correctly to this item. Among students in grades five through seven, approximately as many as agreed that stars are brighter than planets as selected the correct answer; namely, the stars are in the same relative position every night. In these grades there was a slight but not statistically significant difference in favor of the revised filmstrip." (15)

Table 28

Percent Choosing Various Responses to Item 28

Choice of Answer	Grade 5-6		Grade 7		Grade 10	
	FSO	FSRR	FSO	FSRR	FSO	FSRR
N	72	68	72	71	59	61
1	35	38	39	50	71	70
2	14	4	12	8	7	2
3	20	15	7	10	0	3
4	31	43	42	32	22	25

Questions 4, 29, and 28 should definitely have been retested because all student scores were well below criterion. It probably was not necessary to retest question 3 because of the high student score.

Summary

To find out if your revisions were successful you must test the revised unit. If you decide a test is appropriate, then you must determine which tryout elements must be changed for the new circumstances. After you conduct the tryout analyze the results to find out if student achievement meets the desired standard and if a significant improvement is evident.

* * * * *

Recycling, in Brief

Test a revised version of a unit

to determine the quality of changes made and the need for further improvement.

Decide if a test of a revised version is necessary.

Change tryout elements.

When results are in, ask

have students achieved at the level desired, and

has the program improved?

CHAPTER XXIV

The News: Reporting Constructive Evaluation Results

One eminent evaluator said, "The quality of evaluation will not exceed the quality of its communication." (1) One of the most important activities in constructive evaluation is the communication of test results.

CASE

A Constructive Evaluation Report

The following pages contain excerpts from a report to the production staff at the Children's Television Workshop. Is it a good report? What makes it so? After the report each criterion is explained and then applied to this report.

MEMORANDUM

Children's Television Workshop

DATE: January 23, 1973

TO: "Sesame Street" Production

CC:

FROM: "Sesame Street" Research

SUBJECT: Attached Mass of Paper

Dear Production:

Don't despair --- the important parts are in the front, but the fun parts are in the back. The kiddie comments at the end are really worth plowing through --- especially, don't miss Kathy & Claudio, Jimmy, Dennis, & Sadie.

This represents results from a "probe" study on Sam the Machine, Limbo Bits, and Spanish/English bits. Because the study was not of a strict experimental nature, information is heavy in some areas and sparse in others. We have here:

Report on Sam the Machine [a new robot character introduced on Sesame Street]
Report on Limbo Bits [street characters from Sesame Street playing other roles]
Report on Spanish/English Bits [segments shown on a relatively empty set in Spanish and in English
Appendix I - Attention/Distractor Summary for bits
Appendix II - Comments on Miscellaneous Bits
Appendix III - Protocol

SUBJECT: Sam, the Machine Man

Purpose

The purpose of this study was to investigate some children's reactions to Sam, the Machine Man. Fourteen children were shown tapes featuring bits with Sam. After viewing was completed, a researcher talked with the children, prompting their verbal responses to several open-ended questions designed to investigate the following aspects of Sam:

1. Does the child understand Sam's voice?
2. How does the child perceive the reactions of the other cast members toward Sam?
3. How does the child himself feel about Sam?
4. Do the children understand what a machine is?

Comprehension

The children seemed to understand Sam's voice most of the time. Often, however, children expressed difficulty in understanding some phrases or sentences (at one point in the questioning, the tape was stopped as Sam announced, "I hurried over because I heard numbers being spoken." One five-year-old reported this as, "I buried over because I work by smoking.") In some cases, a less garbled machine voice, or less competing background noise (particularly from the machine itself) would do a lot to improve clarity.*

*Outer-Space Cooperation: We tested this bit in order to see if we could generalize about children's comprehension of garbled language (Sam and the Martians being the primary examples of this). What is partially garbled in the audio track is often decoded by the child, who extracts information from the visual track. Therefore, the child's overall impression of the bit is usually correct, but his recall of verbalizations is often incorrect. e.g., the tag at the end - "No, let's call it Shirley" is not understood as a joke. Rather, one child seemed to think that Shirley might be similar to sharing - which are both related to cooperation.

The children reported that Bob and Gordon disliked Sam, and that Oscar liked Sam (because "Oscar wants to be so slick! Oscar wants his clothes on the floor!"). Most of the children themselves reported that they did not like Sam. The children's self-reported dislike of Sam did not adversely affect attention, as the next section of this report indicates.

There seems to be definite confusion about the functions of the machine. Most of the children associated the physical features of the machine with its functions: balloon eyes, sink-drain side, legs which are "shorter than Gordon's." What the children did comprehend about Sam's functions was essentially accurate: that he washes, takes pictures, etc. In general, the children seemed to have little conception about what a machine is and how it differs from a human.

Attention

The following table summarizes distractor data measuring visual attention to the Machine Man bits:

	Age 4	Age 5
1. Machine man does Bob's laundry, Show #424	73%	
2. Bob's laundry is finished, Show #424	86%	88%
3. Machine man finishes Oscar's laundry, Show #424	80%	89%
4. Bob counts 1-10, Show #432	73%	86%
5. Machine man - Gordon & Susan, Show #406	69%	
6. Gordon needs picture taken, Show #447		89% (age 4½)

These bits reflect the overall trend for five-year-olds to have higher attention patterns than four-year-olds. The following seemed to have special attention-pulling power for the children we observed:

The slapstick element of the laundry going on the ground
The noises made by the machine
Physical features of the machine - eyes, gadgets, etc.

Attention seemed to be lowest when Bob and the Machine were arguing (show 432) about whether the machine should count backward. The verbalizations of the machine did not seem especially interesting to the children, who deduced much of the meaning of the episode from the visual track.

Attention rose most dramatically midway through the countdown, and reached a pinnacle as Sam blasted off.

Special Suggestions

The garbled language of the Machine should be made more lucid.

Special features of the machine (blasting off, picture taking, doing laundry) are [illegible]ys very attractive to the children.

A constructive evaluation report must be complete.

Usually evaluation reports are given to producers by evaluators. If a report is incomplete, a producer is likely to make faulty inferences. You can construct a complete report by knowing what a producer wants and what a producer needs to make his decisions. At least you should include the evaluation questions and the details of the four elements mentioned above, tryout procedures, results, comments, explanation of results, and recommended revisions. The description of tryout procedures should tell the whole story about what happened to whom, where, and when. Results should include data, and explanations of charts and graphs. Results should also include incidental unplanned outcomes, and negative and positive findings. Opinions, value judgments, and inferences based on data should be included but should be labeled differently.

CASE

Was it Complete?

The "Sam" report included all the evaluation questions which were asked by Sesame Street researchers, but left out some of the details of the tryout elements and procedures.

331

For example, a producer might want to know who the fourteen children were and how they might be characterized by age, sex, and socioeconomic status. But the selection of subjects may be so standardized at "Sesame Street" that their producer knows that half were boys, half girls, half four year olds, half five year olds, and all of low socioeconomic status.

The "Sesame Street" researchers reported the list of segments tested, but did not report the test site. They did make statements about the measures used. The questions asked of the children were reported in Appendix III. Producers were familiar with the procedure used for measuring attention: the distractor technique.

The evaluators stated the results in two ways; they gave the actual children's responses in the appendix and summarized the comprehension and attention data in the report. The evaluators stated their comments and hypotheses to explain the results found in some cases and not in others; the garbled voice interferes with comprehension; special machine features and functional action by Sam attracts attention, but no hypothesis is given about the children's confusion regarding the functions of the machine. Value judgments and opinions are not given; the evaluators seemed to restrict themselves to data based inferences. The evaluators stated revision recommendations at the end of their report.

A constructive evaluation report should be insightful.

The report should include ideas for improvement and stimulate the reader to think about possibilities and generalizations which could enhance the program's effects. The report should express the ideas in a way which will help the producer decide well and quickly.

If a report includes no insight, a producer is likely to feel annoyed. He may reason that the expenditure of time, energy, and money resulted in no more information or ideas for improvement than one might have made without an evaluation.

You should reveal something not seen by the naked eye. You should show the producer the consequences of each choice of revision to be made.

You should present more than an explanation; you should tell a producer what to do and how better results can be achieved. For example, when you only make statistical statements, you do not tell a producer what to do. You might list by priority the information you gathered and explain how it might be used and with what confidence.

CASE

Was it Insightful?

> In the report on "Sam, the Machine Man" the recommendations made followed from the data. Our recommendation was stated in terms of what a producer should do: the language should be made more lucid. But that recommendation should have been made in the active voice: "The producer should make the voice more lucid." The second recommendation should have been made as a suggestion: "Emphasize the special features of the machine," rather than the generalization "Features are attractive." Each of the suggestions should have included consequences: "To increase comprehension, make the language more lucid." The evaluators might have stated the degree of confidence they placed in their suggestions, "We feel quite sure that these results are valid" or "On a scale of confidence from one to ten we give these results and suggestions a seven."

> A constructive evaluation report should be comprehensible.

No matter what the form of the report may be--written narrative, written or oral question and answers; graph and profiles--the message must be communicated. (2)

The report should be quick and easy to read, see, or hear. You should report on tests that are commonly known and experienced by producers. It should be concise, simple, and stated in the language of producers. Results should include concrete descriptions of student behavior. For example, when Ken O'Bryan reported his eye movement research to "The Electric Company" staff, he showed films of the program which revealed the part of the screen that a student was looking at, by showing a point of light reflected off a child's cornea superimposed on the film.

You should use a few simple labels and concepts and restate them a number of times within the report so that a producer will recognize and be able to interpret tests and methods. The same terms should be used in the same way on successive reports.

If you need technical language, you should define each term. You should be specific. For example, telling someone he should provide appropriate practice is not enough. He must be told that the practice experience should be just like the test experience. Suggestions made in general terms are often misinterpreted. It is easy for someone to believe he's doing something which has been stated ambiguously.

You should present a brief summary of the results before the full report. You should suit charts and graphs to the statistical and arithmetic knowledge of your audience. To be most effective, you should report to a producer personally, face to face. In this way you can detect misunderstandings and rectify them. Never assume that a term used by a staff member has the same meaning as yours. Ask for a definition or example. Avoid jargon. See if the producers get the message by asking them what was said.

CASE

Was It Comprehensible?

The report about "Sam, the Machine Man" was (at least to the producers at "Sesame Street") quick and easy to read. Any technical term in the report (audio track, visual track, tag at the end, distractor) was well known to the producers. Some technical terms unfamiliar to producers, "...verbalizations of the machine...", instead of "statements" or "sounds made by the machine" might have interfered with communication.

The instructional segments used, and the tests referred to, were familiar to the evaluators and the producers. The small table used to summarize attention data was used often and was familiar to the producers. The scores meant something to the producers because they had read numerous reports like this in the past. The memo was accompanied by personal interviews with producers to discuss the results.

A constructive evaluation report must be credible.

You must make the report credible because the information in the report should influence a producer when he makes a decision about program improvement. A report will be credible if you identify and attend to the values and needs of the producer; that is, the report should address significant points as perceived by producers. (3)

Make priorities. Pick the most important things about which to make suggestions and make suggestions which are feasiable within production constraints. If you do not know production limitations, you are more likely to suggest impossible solutions and reduce the chances that a producer will listen to you a second time.

You should review the data for credibility and keep the producer in on the planning. You should use and report tryout procedures if these procedures are perceived as valid methods by producers. To insure that some data is acceptable, you could provide several kinds of evidence and let a producer choose what seems to be believable to him.

The report should be complex enough to accurately represent reality and concrete enough to give a living picture of what happened. The report should have an accurate and correct emphasis. Do not print the report until those who did the evaluation work are satisfied with the accuracy of the statement.

The statements in the report should fit the ethical constraints of a professional society such as the American Psychological Association and include scientific caution and candor. All statements should be supported, and confidential matters should be kept private. Fair comments should be balanced with broad speculations. (4)

There are incidental outcomes to everything we do. The instructional system will have unplanned results. (5) You should state tactfully what's wrong with the program; say it's "not up to standard" rather than it's "lousy." You should include details of materials or methods found not useful or detrimental, and you should state for whom the material is appropriate. (6)

CASE

Was it Credible?

Producers asked the questions listed in the "Sam" report and approved of the tryout procedures used. They were in on the original plans and were informed about the progress of the evaluation. Given their past experience, the producers got a fairly accurate picture of what happened from the report. Protocols which included questions and answers at the end of the report, helped give an accurate account of the tryout.

The suggestions made by the evaluators seem to be substantiated by the evidence they report. They state few opinions. The faults of the segments are tactfully reported and the importance of the faults is not diminished. But the evaluators could have made a statement about the confidence they placed in the results and the suggestions. Generally, the report seems believable.

A constructive evaluation report should be presented quickly.

Depending upon the producers' need for, and their interest in, the information, you may report at any point in the process before or after a tryout. The content of a report may vary, but the criteria are the same for any point in the process.

The message must be communicated to producers quickly, especially in the early stages of the creation of a new segment. If you wait to report, you may find change more difficult, and you may find that you may have to make more than a simple single revision.

You should report to the producer who created the earliest form of the product or method. You should also report to those who have control of the earliest changes if they are different than the producer. The report should be present at the time needed and when the producer is ready to read it. He should have the time to read the report, and he should be beyond the excitement and emotion elicited by the creative stages of producing the unit.

CASE

Was it Presented Quickly?

At the time of the "Sam" report only six segments including "Sam the Machine Man" had been produced. The report came back quickly enough to the producers and writers to put revisions into effect where production costs allowed. But producers have plenty of time to use the suggestions in the creation of new "Sam" segments.

You should try everything possible to see that the report is usable.

One of the main functions of a constructive evaluator is to report back to the instructor or producer. How you report may make considerable difference in the eventual use of the information reported. In performing this critical feedback function, use the following rules.

Generate a procedure to insure the information's reception and use. You cannot sit back and hope that a producer will use the information he receives in a report. You must double check the reception and work out plans to help a producer put the information into practice. You must be prepared to spend time and money to get ideas used.

To communicate from evaluator to producers is harder than communicating within the group of producers or evaluators. So you must help spread the message. Luckily for evaluators, a message does spread. After a message gets through the invisible but existing boundary between evaluation and production sections, it spreads randomly, somewhat like an epidemic. The problem is that the spread is not systematic and predictable. You can make it predictable by checking with each concerned member of production.

Recheck messages: alteration of information always occurs. The amount of distortion depends on the amount of processing which information goes through. Limit the processing so that the message passes directly between you and the producer, face to face.

To translate data to a producer, you must simplify information to make a report understandable. This is sufficient reason to teach the producer the concepts and principles of evaluation or to present raw data and ask for the producers' participation in interpreting the data.

Like explaining the facts of life, don't tell the staff members more than they want to know. Do not try to be overly helpful by suggesting or doing too much; this stifles initiative.

Deliver some criticism indirectly. For example, say, "I wonder how attention can be moved from the picture to the words?" instead of "That picture is distracting."

Do not make direct attacks, even of the mildest sort, on the abilities of the producer. Do not tell him that he cannot judge which instruction is good or bad. Say, "You could make your judgments more accurately if..." or "You could verify your judgments by..." or "You may gain added insights by..."

Evaluation and production staff must be tactful. You can easily alienate subject matter specialists with a thoughtless, "This segment is stupid," or production staff can alienate evaluators with "This report is a waste of paper." Neither can afford to treat the other as merely object or audience; both must deal with each other as people with feelings trying to do a job well.

Never let a producer use information to change the instruction to the point that it will not do what it is supposed to do. In other words, if it's supposed to be instructional, do not let him change it to make it funnier if the instruction will be lost. Negotiate, but not to the extent that negotiation damages the objectives of the instruction. If you do agree to some changes which run counter to the intent, and the instruction fails to do all you said it would, your rationalizations will sound like sour grapes.

Reports to producers should be delivered in their preferred mode. Some prefer charts and graphs; some like raw data, such as verbatim quotes; some prefer the information in writing, some in conferences. The method of reporting should be similar to the kind of feedback they might get in ordinary occasions. (7) For example, a T.V. producer likes to monitor people's reactions as they are exposed to his product. Therefore the reporting technique should show a producer a film of viewers' reactions.

To encourage the use of information collected, constructive evaluation procedures and results, all staff members must know and reach agreement on objectives, target population, and procedures.

When a modification is in order, inform all those people who have responsibilities related to production of the changes that will be required. Each change has many effects -- not all foreseeable -- and all concerned must know about the change. When a change is suggested in a script, propmen, stage hands, actors, producers, and cameramen have to know about it to make the revision efficiently.

Convert the results into a growing list of suggestions and teach the producers how to use the suggestions. Check to see if they follow through with valid revision.

Try to predict the reactions of producers with different personalities to your comments. You may know that someone is sensitive about the humor in his segments, or is terribly excited about one particular creation; in that case you may want to soften or postpone your comment.

Let him draw his own conclusions. Be alert to your own motives and to the producer's motives. Some evaluators feel that to make research credible they have to contradict producers' hunches or confirm their own ideas. Producers are likely to accept results when it reinforces their thoughts, when it is presented tactfully, and when suggestions for improvement are included.

When suggesting a change, do not make the modification a point of challenge or of win or loss: (8) set the emotional climate so that neither evaluator wins or loses when a result reveals the need for revision. (9) No one likes to be told he was wrong and that he has to redo something in which he invested his pride and lost, and no creator should be made to feel that he is no good because a first draft of his work was not effective.

If time permits, make any necessary changes gradually. Have the producer put into practice small parts of a major change or a mini-version of a major change that will bring maximum learning.

Summary

An accurate, understandable, acceptable report is one necessary step to produce instructional improvement. Without communication between evaluator and producer not even the best measures and results will save an instructional method.

* * * * *

Reporting Constructive Evaluation Results, in Brief

A good report is

- complete
- insightful
- comprehensible
- credible
- presented quickly
- usable.

CHAPTER XXV

The Odd Couple: Working Toward Commitment

In the process of instructional development, commitment refers to any behavior which can be described as seeking to improve instruction. Thus, when you request information on how to improve a project, and you use that information to make changes, you are demonstrating commitment.

In the development of an institutional project, each staff member must have the desire to improve. In the development of large-scale instructional projects there are usually some people given the responsibility of creation, and others the responsibility of evaluation. The project effort will have been for nothing, and there will be little improvement, if the creators do not accept and use the information gathered by the evaluators. But there is a natural antagonism between those who produce instructional methods and materials and those who seek to improve what is produced. No one really wants others to find fault with their work, and no one wants to revise what they thought was an adequate product. Yet there are few projects which turn out to be effective, efficient, and acceptable on first draft. If a director wants to create an effective project, it must be improved, and to improve, those who produce the instructional methods and those who find the strengths and weaknesses of the instructional methods must cooperate. A project director must plan carefully to achieve the degree of cooperation needed between a creator and an evaluator, the odd couple.

A good indication of a producer's attitude toward revisions is the speed with which he puts revisions into effect. The differences between producers are great. One producer may take a day to begin work on changes, another three months; another may never consider revision. (1)

A producer is not likely to make revisions from reading an evaluation report; he must first be committed to improvement. A producer must show commitment by giving time and money for revision. (2)

CASE

Demonstrating Commitment

Producers and evaluators at the Children's Television Workshop (C.T.W.) -- the creators of "Sesame Street" and "The Electric Company" -- are committed to improvement.

Changes are continually being made on the basis of research findings. For example, when placement and movement of print on the screen were found by researchers to influence the movement of children's eyes, the producers used what the researchers had found. The movement of print, and direction of a character's actions toward letters and words, were taken into account to make sure that children would see and scan the words on the television screen.

As another example, consider that a confusing segment was changed by a writer on the basis of a researcher's comment about a script: to teach enumeration (counting objects 1, 2, 3, 4), a writer planned to show four dice, each respectively showing one, two, three, and four dots with the numeral appearing above each die. A researcher pointed out that the four dice should all have had the same number of dots, or that the segment should have included one die with four dots, so as not to cause confusion between the number of dots and the number of dice being counted.

The staff members' wish to work together to improve was present early in the formation of the workshop, and the staff's attitude was evidenced in this remark:

> "One of the many achievements of the Workshop has been the successful fusion of production, professional education, and research." (3)

The personality, views, and habits of each staff member, and the structure and workings of the organization in which the instructional project is being developed, contribute to the working relationship of the staff and their commitment to improvement.

The factors which contribute to a successful cooperative working relationship among people of different viewpoints are those which also influence a successful marriage: much depends on the views and habits of each person, but the stresses and strains of the moment also make an impact on the relationship. These factors should be taken into account when you promote a cooperative relationship to improve an instructional project.

To maintain an optimal productive relationship among staff members working on an instructional project, and to promote commitment to improve, each person should be sure of his role in the cooperative endeavor.

Each person's role and responsibilities should be spelled out: each should know how much he controls of production, budget, curriculum, testing, scheduling, and writing. Everyone should know the responsibilities and the roles of other staff members, and who has the authority to make the final decisions.

When there is division of labor in a large project, some people may be designated as evaluators and others may be considered as producers. An evaluator's role may vary, from that of an independent outside authority with no special commitment to the project, to the role of an involved full-time team member with complete knowledge of the project.

An evaluator should serve a producer, and a producer should create the methods and materials. A producer should come to an evaluator with questions; an evaluator should help the producer answer the questions. A producer must know the quality of his creative efforts, an evaluator should provide useful evidence of the strengths and weaknesses of a method or product, and allow the producer to use this evidence in making his own decisions. A producer should make production decisions; an evaluator should make suggestions, not production decisions. In most cases an evaluator should leave the producer free to decide what will be done to the instructional method. (4)

An evaluator should check to see how a producer's work is going and how a project is progressing. At the beginning he should explain that he will be observing in order to give feedback and thus add precision to the producer's techniques. He should explain that he is not snooping or trying to threaten. He should not check or make demands before a producer is ready, for the producer may be embarrassed. Instead, a producer and an evaluator should make up a mutually satisfactory schedule of appointments.

A producer should remain open to questions suggested by his evaluators. But an evaluator should tell a producer at the beginning to expect the cyclical and continuing process of revision; otherwise, producers may operate under the assumption that one test of a product will be all that is necessary.

A producer should make use of the information collected. An evaluator should encourage a producer to use constructive evaluation. But this is easier said than done. The discovery of faults and weaknesses, the primary results of constructive evaluation, hurt a producer no matter how well prepared he is to receive the news. The best an evaluator can do is point out the positive results first, give praise for doing evaluation and for any ideas suggested for revision. When revisions are fruitful, he should praise the producer for insights gained from constructive evaluation.

When giving bad news, an evaluator should prepare a producer. He should explain that there are always negative results, that bad news is what they are looking for, and that there are reasons for looking for it. He should stress that he is trying to help and to add precision to what the producer already does well. An evaluator should have positive suggestions ready if it appears that the producer will be completely at a loss as to what to do next.

An evaluator should reward any attempts on a producer's part to make changes. He should reward risk-taking and the willingness to try new things, even when mistakes are likely to occur, and he should make sure that the producer is getting some results for what he attempts to do. He should help a producer to use pieces of his new knowledge immediately. He should have the producer experiment, and then, contingent upon the resulting evidence, spend time with him, and give him some helpful suggestions.

In a large project, it is advisable for a director to appoint some person or people to act as go-between for evaluators and producers. The liaison should know the most recent research information, have enough time to watch producers create the projects, review instructional plans and drafts, suggest changes, discuss research results with producers, and see that plans are accurately translated into the final product. The liaison should know who is responsible for each production task so that any problem can be brought directly to the person who can solve it.

To do all this well a liaison needs the trust and respect of the producer or instructor. The producer must be confident that the liaison will not let poor work slip through or be dishonest in his criticism.

CASE

Defining Roles

> The original C.T.W. team was small: each member knew what his role was. The production section was to get the show out; research was to help production make the best possible show. To do this, researchers collected data on the show's appeal and the show's effectiveness. Researchers continually recorded examples and teaching strategies in a writer's notebook, from which writers selected ideas for sketches on "Sesame Street" which would lead to learning. Researchers reviewed scripts to check the show's ability to teach. Researchers also watched the studio action as videotaping took place and provided advice to production staff when educational aspects of the performance could be improved.
>
> Producers and writers created the show. They also listened to researchers and learned educational principles to be used to achieve the effects they wanted to produce. (5)

> To maintain an optimal productive relationship and to promote commitment to improve, each person should be confident enough of his own abilities and skills to be able to risk asking questions and risk making decisions based on sources which contradict his intuition. He must be open to changes made in his creations, and to views other than his own.

Each staff member must be chosen for his ability *and* his confidence in his ability. In other words, each person should know what he knows and be willing to ask questions about what he does now know.

Producers who lack confidence in themselves may rationalize that constructive evaluation will inhibit their creativity. But evaluation can be a catalyst for creativity. Results can provide the stimulus to break through rigid assumptions and open new boundaries. For example, Kenneth O'Bryan, a researcher, demonstrated to producers of "The Electric Company," a television show designed to teach reading, that a child's eyes do not readily scan words placed at the bottom of the screen: that they should feel free to break with this dominant approach. (6)

CASE

Choosing Confident Staff

> Each team member at C.T.W. knew that he was picked because he was well-qualified in his area. The professional T.V. producers were not expected to know about education, and educators were not supposed to know television production. Because staff members were sure of their own abilities, and their knowledge in some areas was expected to be limited, a free exchange of questions and information took place. It was not difficult for them to realize that an

esthetically pleasing T.V. production might not necessarily be a sufficient experience to get a child to learn to read. As one of them later commented:

> "The television professionals were unconcerned about their academic egos, since they had none to protect, and therefore felt unconstrained: we were not afraid to ask the dumbest questions in the world, because we were not expected to know anything about these kids." (7)

There is a great deal of give and take between producers, writers, and researchers. As this text is being written, as the following examples show, researchers and producers at C.T.W. are still cooperating to find out how effective "Sesame Street" and "The Electric Company" are and what changes to make.

A producer approaches a researcher to ask him to find the best way to put print on the screen so children will read it; to find out if some new segments, such as those including a robot called Sam the Machine, are appealing to children; to find out if new goals are too hard for children to achieve; or to find out if certain segments are teaching children to solve problems. Writers meet with researchers and discuss how best to reach a goal. A writer asks a researcher for examples of a consonant blend, or for ways to put print on the screen, or for a series of rhyming words, or for methods of accentuating parts of a word. A film staff member brings animation storyboards to a researcher to see if educational principles are being used or violated. And researchers ask producers what techniques the producers feel are contributing to attention and comprehension of show material so that the right kind of research questions can be asked.

Each person should be able to compromise in a conflict situation.

In most cases evaluators should not challenge producers and vice versa. An evaluator should not use evaluation in a personal vendetta, to prove an evaluator's point or to show a producer there is a mistake in his instructional design. When a challenge is made, a liaison person should present the problem to producers and researchers. If both groups include secure, confident people they should be able to compromise.

CASE

Compromising

> The cooperative relationship between researchers and producers at C.T.W. is not perfect. There seems to be a healthy tension between producers and researchers which promotes a continual reexamination of the function of research and its usefulness. Occasionally, a producer or writer is annoyed by the results of an evaluation and considers the results an insult. When producers don't follow advice given by evaluators, the evaluation staff is sometimes insulted. Luckily, there are some sensitive staff members who can communicate with both parties. These researchers communicate the functional relationship between production and research staff so that the two departments can work together to produce the best show.

Relations among staff members and commitment to improve will be enhanced if the organization in which they work provides a goal or purpose for a project which is of high priority among the values of a producer and evaluator.

All team members must talk to each other about the goals, the system, and the process of development. They should arrive at an agreement about their intentions. If the intentions of the group correspond to the values and aspirations of each individual, the group will function well and will want to improve its work.

CASE

Choosing Goals Corresponding to Staff Values

> The original team working on "Sesame Street" was not concerned about status among producers and evaluators. Their eyes were all focused on what they felt was an important societal need: the education of culturally deprived children. (8)

The organization should make a special effort to foster cooperative relationships.

Don't add to your problems by antagonizing staff members. If you decide to pursue a constructive evaluation strategy, institute it gradually: a quick dose of critical evidence can be rough on a producer. The typical producer's reaction to information collected about his product is hardly in the same category as an infant's confused perception at birth, but it is sometimes painful, often surprising and shocking. The effect is magnified if the existing instructional system has been in use for some time.

Don't frighten staff members away. Do not make evaluation demands too early in the process. In the beginning, deal with a team leader only; hold back from making demands of the rest of the staff until some substantial progress is being made.

See to it that any interaction relating to constructive evaluation is pleasant and easy. Make contacts brief; ensure there is no fatigue and that enjoyment of these encounters persists. Make the encounters productive and task-oriented.

Trust is an essential feature of a collaborative effort. When trust is established among members of a small team which is charged with accomplishing a challenging task, ideas are more readily expressed and more honestly accepted or rejected. You can gain trust by helping a producer achieve the instructional goals, by keeping promises, holding lines of communication open, and otherwise doing anything that shows you care about the effort.

CASE

Fostering Cooperation

The partnership of producers and evaluators at C.T.W. was carefully planned during the first collaborative effort, a seminar to determine goals. The seminar had a specific focus: social, moral, and affective development; language and reading; mathematical and numerical skills; reasoning and problem solving; and perception. Researchers, educators, artists, children's authors, entertainers, teachers. C.T.W. staff and sponsoring representatives attended. Issues were identified in advance and short papers were prepared on topics to orient the meeting.

Each meeting was run precisely. Joan Cooney, President of C.T.W., provided guidelines and purpose: the show had to be entertaining, it had to appeal to older children to get them to tune in to the program, and the program had to teach without the aid of teachers and books. A psychologist then explained what 4-year-olds could learn. Prepared comments were read, the goals suggested, and discussion followed each paper.

Notes were organized, typed and distributed by the morning of the second day. Small groups were formed by the chairman, Dr. Lesser, to encourage the greatest possible participation when discussing promising topics consolidated from the notes of the previous day. The second day's meetings were the most productive. The third day consisted of group reports.

The precise planning and effort to make educators and producers work well together was shown in several ways:

During the conference professional educators often lapsed into jargon and technical terminology which created a barrier between themselves and producers; the C.T.W. staff struggled to tear the barriers down:

> "On these occasions the staff seemed to take on the characteristics of a Greek chorus, intoning repeatedly, 'What do you mean by that?' What do you mean by that?' This continued until adequate, simple explanation would be forthcoming...These conditions clearly prevented technical discussions from spinning off into the stratosphere, with people believing or pretending that they understood each others' language and frames of reference, but not really doing so." (9)

By compromising on the approach to educational problems and by sticking to the task, the fundamental conflict between producers and evaluators became apparent: production experts felt that creating a program is based on intuition; educators felt that a program could be designed deliberately and systematically:

> "They contended that any book, film, music, or television program -- indeed all creative products -- can only be conceived intuitively and lovingly, with the creator drawing freely upon his own fantasies, feeling, and experiences; the dissection of deliberate thought and methodical planned analysis destroyed the naturalness that must be inherent in the product." (10)

Yet through the guidance of group leaders a compromise was reached.

> "Temporary armistices usually took this form: academics and educators -- presumably the thinkers and analyzers -- acknowledged the necessity of intuition in designing creative materials but argued that adding some elements of analysis in deliberate planning need not smother that necessary intuition. The protesters were skeptical of this compromise, but they also were eager to avoid a stalemate. They agreed that since we were meeting to exchange thoughts about the goals of a children's television series, we should proceed in the unlikely hope that thought and intuition were not inevitably incompatible. No one really was convinced, but the confrontation usually ran its course in this way and then everyone went back to the work of redefining the goals for the series." (11)

By selecting flexible participants, the C.T.W. staff conducted a conference of diverse personalities and points of view and encouraged a great deal of give and take:

> "A few observations were common to all participants no matter what their professional background. Everyone needed to break old habits of thought and apply himself with agility to a task without precedents. All needed to suppress practiced speeches designed to display cleverness and elegance of phrasing. Everyone needed to avoid punishing other participants verbally and to meet confrontations with humor and flexibility. With the constant risk of fragmented, non-consecutive conversation in a large group, everyone had to adapt his behavior to avoid this. All needed to

> listen, and this required stamina. All needed to contribute to a momentum, an energy and liveliness that would keep the sessions moving ahead. Many succeeded and added greatly to the project's chances; some did not." (12)

By using tact, those individuals who went beyond the limits of the conference or provoked hostility and blocked progress were handled:

> "By convincing people that in one way or another he liked and respected them, Lesser, later in the sessions, was able to indicate to an individual that he was 'out of line,' dealing on a false issue, or unnecessarily expanding a topic without that person feeling great amounts of hostility or embarrassment. If hostility was aroused and perceived, Lesser would attempt to allay these feelings during a conference break."
>
> "If a person needed to be redirected (or effectively shut up), he either did not understand the ground rules, had missed a point about the purpose of the seminar, or suffered from some other sort of momentary confusion." (13)

The result of the conference was that producers understood goals and felt as if the goals had not been imposed on them.

From day to day, starting with the first conference and the first tests of the pilot, the research department worked at maintaining a cooperative arrangement with production. People at the workshop recognized that the relationship had to be worked at.

> "You not only have to do research, but you also have to make it appealing. You have to communicate it in ways that are understood and liked. You have to play politician while doing research and be diplomatic about it. Research is not there to tell the producers what to do. It is they who are responsible for turning the last crank. You can't look over their shoulders too closely, or you make yourself obnoxious."
>
> "...If the research didn't deserve the audience of the producer, probably it wasn't speaking to his problems..." (14)

> "I always felt that the producer should participate in the research from before the time it's done. I can bring in research results as end-point conclusions from research projects, and I can lay them on the producers' desks. They will be courteous about it. They will read it. They're nice guys. But I involve the producers in the initial design of the study, let them review my plan just before it goes out into the field and make suggestions for revisions and extensions. Then they are sitting there waiting eagerly for the results to come in, and sometimes they have their shirtsleeves rolled up helping you plot the data. Moreover, we take them out to the field so that they see the methods and procedures in use. This way they develop a hands-on sense of what the study is all about and actually see how the children are responding, instead of having to see only field researchers' written reports." (15)

Researchers showed their concern for producer's efforts early. When a researcher would overhear a conversation or be asked a casual question, he would follow it up with an answer some time later. When it was apparent that research could provide answers, production started to ask questions.

> "There was, for example, the question of whether it was feasible to use the spot-announcement technique for instruction, based on the element of repetition. Would all types of materials bear up under repetition? Would some bear up better than others, less than others? It is important to find out what does not work, as well as what does work. Would the youngster continue to watch the commercial? Would he pick up jingles? Would he learn more from listening once? Is it possible to build a kind of hierarchy sequence of instruction within a one-minute segment, so that the child learns something the first time he sees it, adds something the next time, and so forth?" (16)

The willingness of producers to improve their work based on constructive evaluation was one of the factors leading to the ultimate success of "Sesame Street." The producer-researcher relationship undoubtedly contributed to the commitment observed.

The organization should give control and freedom to each person at his own level of responsibility and ability and make each person feel that he is contributing.

CASE

Giving Control and Freedom

How was the relationship between production and research at C.T.W. built? The major forces behind the formation of the Workshop took into account many of the factors mentioned which influence commitment.

The original staff gave complete control of the creative endeavor to the production department. The producers and writers did not have to accept suggestions for teaching strategies or teaching goals from administrators or researchers. Consider these quotes:

> "That was a vast change in educational television--in that the bosses were the entertainers, not the educators." (17)

> "What the Workshop management has grasped is the importance of involving it [evaluation] in the building phases from the beginning, and of doing it in such a way that they genuinely feel they have full creative control. This is seen in the care with which the job of setting goal priorities was approached, keeping in mind that the staff had already participated in the preliminary adventure of the seminars."

> "One of the reasons I've been happy here is because Jon (Stone) and I, and the other people who put it together in the beginning were left absolutely 100 percent alone. There were no sponsors looking over our backs. Joan Cooney wasn't looking over our backs. I'd say that in two full seasons of "Sesame Street," Joan Cooney has made two comments to us about either do this or don't do this on the show. We were left alone. She said: 'Put on a television show.' She knew she had the people to do it." (18)

> In some cases the advice is being taken, in others, the production staff is not using the advice.
>
> "When you produce a show, you're exposing yourself to the world...we were scared enough at that point, I think, so that we wanted all the help we could get. It's the overall attitude of the operation. We don't have to do anything these people tell us. We can do precisely what we want to do--but let's hear what they have to say about it. In some cases, people made suggestions that we ignored. So you have a little confidence to perhaps overcome that exposure factor, if you know that you can say, 'Well, I think he's crazy.'" (19)

An evaluator can do everything well and still feel that he has failed because the information he collected is not used. One of the major reasons a producer fails to use information is the lack of time, money, and staff to do so. You, as project director, should make sure that a producer has enough resources to carry out ideas inspired by the information provided.

CASE

Keeping Open to Change

> "As Connell notes, the premiere of the program on November 10, 1969, marked a stepping-stone rather than an end-point to the research-production cooperation. Throughout the period of the telecasts, formative research studies continued to guide the development of new production techniques, format elements and teaching strategies. And the research goes on, reflected in the ceaseless effort of the producers to improve the program." (20)
>
> "To appreciate the historic nature of what occurred, it is necessary to understand that the C.T.W. was quite prepared to scrap all five hours of programming completely if they failed to live up to expectations as measured by the tests, an unheard of practice in television when an out-of-pocket investment of $230,000--the actual expenditure--is involved." (21)

> "Teaching young children by television must be considered a self-correcting experiment: therefore, its curriculum must remain open and flexible to allow changes in response to information as it accumulates. The early versions of a curriculum for television inevitably will include certain objectives that turn out to be inappropriate for televised teaching and will exclude some of great potential value. In the absence of good evidence, these early efforts to construct a curriculum will underestimate certain skills of pre-schoolers and overestimate others, and must be adjusted and refined through successive approximations based on observations of children as the limits of the medium are tested." (22)

> "The unique aspects of this operation are the research aspects. It is no accident that the show is a blockbuster. It was researched within an inch of its life. We knew for a fact, when we went on the air, that the pieces we had in the show would test out very high. We really didn't know it was going to become the hit that it is. But a year and a half of our careful research had gone into this. I would recommend it as an absolute must to anybody who is putting together a television experiment." (23)

Summary

You have to work to motivate the odd couple to work together. The individuals and the organizations have to do everything possible to encourage people to work together to improve instructional projects.

* * * * *

Working Toward Commitment, in Brief

To produce commitment to improve instruction

- define roles.
- choose confident staff members.
- choose open personalities.
- arrange compromises.
- provide goals compatible with the values of staff members.
- foster cooperative relationships.

give control and freedom at certain levels of responsibility and ability.

make each feel he is contributing.

keep the project open to change.

References

Chapter II

(1) Markle, D. G., *Final Report: The Development of the Bell System First Aid and Personal Safety Course, An Exercise in the Application of Empirical Methods to Instructional System Design*, American Institutes for Research, Palo Alto, 1967.

Chapter III

(1) Baker, L., Schutz, E., *Instructional Product Development*, Van Nostrand Reinhold Co., New York, 1971.

(2) Rosen, M. J., *An Experimental Design for Comparing the Effects of Instructional Media*, Ed.D. Thesis, University of California, Los Angeles, 1968.

(3) Morris, H. J., Wallace, H. M., *Programmed Instruction and Hospital Training Problems*, Ann Arbor, Michigan, November, 1968.

(4) Merrill, I. R., McAshan, H. H., "Predicting Learning, Attitude, Shift, and Skill Improvement from a Traffic Safety Film," *AV Communication Review*, vol. 8, pp. 263-74 (1960).

(5) Hovland, C. I., Lumsdaine, A. A., and Sheffield, F. D., *Experiments on Mass Communication*, Princeton University Press, Princeton, 1949.

(6) Twyford, L., "Film Profiles," *Instructional Films Research Reports* (SDC 269-7-23), U.S. Naval Special Devices Center, Port Washington, 1951.

(7) Robeck, M., *A Study of the Revision Process in Programmed Instruction*, M.A. Thesis, UCLA, Los Angeles, 1965.

(8) Weil, J., "Research and Evaluation Section, The Health Show Proposal," First Draft, Children's Television Workshop, New York, October 16, 1972.

(9) Stake, R. E., "Objectives, Priorities, and Other Judgment Data," *Review of Educational Research*, vol. 40, no. 2 (April, 1970).

(10) Dichter, E., *Handbook of Consumer Motivations*, McGraw-Hill Book Co., New York, 1964.

(11) Haney, J. B., Lange, P. C., and Barson, J., "The Heuristic Dimension of Instructional Development," *AV Communication Review*, vol. 16, no. 4 (Winter, 1968).

(12) Sanders, J. R., Cunningham, D. J., *A Structure for Formative Evaluation in Product Development*, Educational Research and Evaluation Laboratory, Indiana University, Bloomington, March, 1972.

(13) Weil, J., "Health Show Prospectus," Children's Television Workshop, New York, 1972.

(14) Sedlik, J. M., *Systems Engineering of Education XIV: Systems Techniques for Pretesting Mediated Instructional Materials*, Education and Training Consultants Co., Los Angeles, 1971.

(15) Wolff, C. A., *Cumulative Effects of Three Major Emphases on Films on Attitudes of Basic Trainees*, Bureau of Social Science Research, Washington, D.C., January, 1964.

(16) Stephens, J., *The Process of Schooling, A Psychological Examination*, Holt, Rinehart & Winston, New York, 1966.

(17) Dubin, R., Taveggia, T. C., *The Teaching-Learning Paradox*, University of Oregon, Eugene, 1968.

(18) Weil, J., "Research and Evaluation Section, The Health Show Proposal," First Draft, Children's Television Workshop, New York, October 16, 1972.

(19) Paulson, C. F., "Evaluation of Instructional Systems," Section IV, *National Research Training Manual*, Edited by Jack Crawford, Teaching Research, Division of the Oregon State System of Higher Education, Monmouth, 1969.

(20) Stake, R.E., "Objectives, Priorities, and Other Judgment Data," *Review of Educational Research*, vol. 40, no. 2 (April, 1970).

(21) Stake, R. E., "Language, Rationality, and Assessment," *Improving Educational Assessment and An Inventory of Measures of Affective Behavior*, National Education Association, Washington, D.C., 1969.

(22) Weil, *op.cit*.

(23) Sanders, J. R., *Comments on Four Case Studies of Formative Evaluation*, Educational Research and Evaluation Laboratory, Indiana University, Bloomington, 1971.

(24) MacMillan, C. J. B., McClellan, J. E., "Can and Should Means-Ends Reasoning Be Used in Teaching?," *Concepts of Teaching: Philosophical Essays*, Edited by C. J. B. MacMillan and Thomas W. Nelson, Rand McNally, Chicago, 1968, pp. 119-150.

(25) Weil, J., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972.

(26) Weil, J., "Health Show, Working Proposal," Children's Television Workshop, New York, 1972.

(27) *Ibid*.

(28) "The Sesame Street Writers' Notebook IV," Children's Television Workshop, New York, 1971.

(29) Excerpt from Pre-script No. 2 - "The Market Place," Bilingual Children's Television, Oakland, California, 1972.

(30) Sullivan, H. J., Baker, R. L., and Schutz, R. E., "Developing Instructional Specifications," in Baker, R. L. and Schutz, R. E., *Instructional Product Development*, Van Nostrand Reinhold Co., New York, 1971.

(31) Weil, J., "In House Health Show Memo," Children's Television Workshop, New York, 1973.

(32) Abedor, A. J., *Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems*, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(33) Sedlik, *op.cit*.

(34) Grasha, A. F., *Evaluating Teaching: Some Problems*, Institute for Research and Training in Higher Education, University of Cincinnati, vol. 4, no. 3 (Spring, 1972).

(35) Sedlik, *op.cit*.

(36) Miller, J. G., "Toward a General Theory for the Behavioral Sciences," *The American Psychologist*, vol. 10, no. 9 (September, 1955).

(37) Abt, C., "How to Evaluate the Cost-Effectiveness of Games," Chapter 8, *Serious Games*, The Viking Press, New York, 1970.

(38) *Ibid*.

(39) Seiller III, K., *Introduction to Systems Cost-Effectiveness*, John Wiley and Sons, New York, 1969.

(40) Rosen, *op.cit*.

(41) Alkin, M. C., *Evaluating the Cost-Effectiveness of Instructional Programs*, Center for the Study of Evaluation, Los Angeles, no. 25, May, 1969.

(42) Scott, R. O., As Interviewed by Stephen Yelon, New York, 1972.

(43) Seiller, *op.cit*.

(44) Baker, R. L., Schutz, R. E., *Instructional Product Development*, Van Nostrand Reinhold Company, New York, 1971.

(45) Popham, J. W., Baker, E. L., *Rules for the Development of Instructional Products*, Southwest Regional Laboratory for Educational Research and Development, Inglewood, 1967.

(46) Daniel Yankelovich Incorporated, *A Report on the Role and Penetration of Sesame Street in Ghetto Communities* (Bedford Stuyvesant, E. Harlem, Chicago, and Washington, D.C.), Prepared for Children's Television Workshop, New York, April, 1973.

(47) Scott, R. O., As Interviewed by Stephen Yelon, July, 1972.

(48) Haney, *op.cit*.

(49) Weil, J., "Health Show Proposal," Children's Television Workshop, New York, 1972.

(50) Abedor, *op.cit*.

(51) Sedlik, *op.cit*.

(52) Allen, W. H., "Audio-Visual Communication," *Encyclopedia of Educational Research*, Edited by Chester Harris, The MacMillan Co., New York, 1960.

(53) Hoban, C. F., Van Ormer, E. B., *Instructional Film Research, 1918-1950*, Technical Report No. SDC 269-7-19, U.S. Naval Training Devices Center, Port Washington, December, 1950.

(54) Special Devices Center, *Instructional Film Research Reports: Vol. I*, Technical Report No. SDC 269-7-36, U.S. Naval Training Devices Center, Port Washington, January, 1953.

(55) Special Devices Center, *Instructional Film Research Reports: Vol. II*, Technical Report No. SDC 269-7-61, U.S. Naval Training Devices Center, Port Washington, June, 1956.

(56) Melaragno, R. J., Newmark, G., "A Pilot Study to Apply Evaluation Revision Procedures in First-Grade Mexican-American Classrooms," System Development Corporation, Santa Monica, 1968, p. 17.

(57) Weil, J., "Health Show Proposal," Children's Television Workshop, New York, 1972.

(58) Mielke, K., "Memorandum," Children's Television Workshop, New York, 1973, pp. 4-5.

(59) *Ibid*., p. 2.

(60) *Ibid*., p. 3.

(61) Brown, L., "A 'Sesame Street' for Adults on Health Care Tested," *The New York Times*, November 12, 1973.

Chapter IV

(1) "The Electric Company" Staff, As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972-1973.

(2) *Ibid.*
(3) *Ibid.*

Chapter V

(1) Chen, M., "Verbal Response to 'The Electric Company': Qualities of Program Material and the Viewing Conditions Which Affect Verbalization," The Children's Television Workshop, New York, 1972, pp. 4-5.

(2) Castrup, J., Scott, R., and Ain, E., "The SWRL Kindergarten Art Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, May 18, 1970, p. 10.

(3) Bilingual Children's Television, Research Division, *Curriculum*, Oakland, California, 1972.

(4) Walker, J., As Interviewed by Stephen Yelon concerning Hood, et. al., *Design of a Functional Competence Training Program for Development, Dissemination and Evaluation Personnel at Professional and Paraprofessional Levels in Education*, Far West Laboratory for Educational Research and Development, Berkeley, 1973.

(5) Castrup, *op.cit.*, p. 9.

(6) Piper, R., As Interviewed by Stephen Yelon, Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, 1972-1973.

(7) Bernstein, L., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, January, 1973.

(8) Waks, S., *The Development and Testing of an Instructional Model for Laboratory Experiments on Electronic Circuits in College-Level Engineering*, Ph.D. Thesis, Michigan State University, East Lansing, 1973.

Chapter VI

(1) Grass, R. C., "Disciplined Creativity: The Latest Shotgun Wedding," The Association of National Advertisers, Inc., New York, 1968.

(2) *Ibid.*

(3) Sedlik, J. M., *Systems Engineering of Education XIV: Systems Techniques for Pretesting Mediated Instructional Materials*, Education and Training Consultants Co., Los Angeles, 1971.

(4) Sullivan, H. J., "Product Development Documentation and Review Guidelines," Southwest Regional Laboratory for Educational Research and Development, Inglewood, California, January, 1968.

(5) Grass, *op.cit.*, pp. 7-8.

(6) *Ibid.*, p. 4.

(7) Nimnicht, G. P. Rayder, N. F., and Alward, K., "An Evaluation of Nine Toys and Accompanying Learning Episodes in the Responsive Model Parent/Child Component," Occasional Research Report No. 5, Far West Laboratory for Educational Research and Development, Berkeley, June, 1970, pp. 4-5.

(8) "The Electric Company" staff, As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972-1973.

(9) Emmer, E. T., Millett, G. B., *Improving Teaching Through Experimentation, A Laboratory Approach*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1970.

(10) Scott, R. O., As Interviewed by Stephen Yelon, July, 1972.

(11) Faison, E., Rose, N., and Podell, J. E.,"The Effects of Test Pauses During Training Film Instruction," 1955, Summarized in *Student Response in Programmed Instruction*, Edited by Arthur A. Lumsdaine, National Academy of Sciences - National Research Council, Washington, D.C., 1961.

(12) *Ibid*.

(13) Alexander, L. T., Davis, R. H., "Developing a System Training Program for Graduate Teaching Assistants," ESSO Education Foundation Grant, Michigan State University, East Lansing, February, 1970.

Chapter VIII

(1) Sedlik, J. M., *Systems Engineering of Education XIV: Systems Techniques For Pretesting Mediated Instructional Materials*, Education and Training Consultants Co., Los Angeles, 1971.

(2) Bryant, Jr., J., "Measures of Attention for C.T.W. Programs," Children's Television Workshop Memorandum, New York, 1972.

(3) Sedlik, *op.cit*.

(4) Calder, J. R., "Attitude: The Unexplored Dimension in Teaching," *NSPI Journal*, pp. 5-10 (December, 1970).

(5) Becker, S. L., "Reaction Profiles: Studies of Methodology," *Journal of Broadcasting*, vol. 4, pp. 253-268 (Summer, 1960).

(6) Becker, S. L., *The Relationships of Interest and Attention to Retention and Attitude Change*, University of Iowa Press, Iowa City, Iowa, 1963.

(7) Mielke, K. W., Bryant, Jr., J., "Formative Research in Comprehension of CTW Programs," Children's Television Workshop, New York, June 30, 1972.

(8) Grass, R. C., Winters, L. C., and Wallace, H. W., "A Behavioral Pretest of Print Advertising," *Journal of Advertising Research*, vol. 2, no. 5 (October, 1971).

(9) Mielke, K. W., "Recording the Amount of Attention Given to Various Portions of the Screen," Children's Television Workshop, New York, July 27, 1972.

(10) Weil, J., "Research and Evaluation Section, the Health Show Proposal," Second Draft, Children's Television Workshop, New York, October 24, 1972.

(11) *Ibid*.

(12) Horner, V. M., "Some Preliminary Notes on Ken O'Bryan's Eye-Movement Studies of 1972/3 Segments," Children's Television Workshop Memorandum, New York, February 21, 1973.

(13) Weil, *op.cit*.

(14) Reeves, B., "Report of Research on First Five Sesame Street Shows," Children's Television Workshop, New York, September 24, 1969.

Chapter IX

(1) Bryant, Jr., J., "Measures of Attention for C.T.W. Programs," Children's Television Workshop Memorandum, New York, 1972.

(2) Scott, R. O., "Program Development Through Formative Evaluation: The SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, August, 1970.

(3) Locatis, C., Smith, F., "Guidelines for Developing Instructional Products," Educational Technology (April, 1972).

Chapter X

(1) Palmer, C. A., "Commercial Practices in Audience Analysis," Journal of the University Film Producer's Association, vol. 6, pp. 9-10 (Spring, 1954).

(2) O'Connor, P., "The Use of Course-Specific Questionnaires in Formative Evaluation," University of Michigan, Ann Arbor, A paper presented at the AERA Convention, 1973, p. 3.

(3) Ibid., p. 4.

(4) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

Chapter XI

(1) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

Chapter XII

(1) Gropper, G. L., "Does Programmed Television Need Active Responding?," AV Communication Review, vol. 15, pp. 5-22 (Spring, 1967).

(2) Flanagan, J. C., "The Uses of Educational Evaluation in the Development of Programs, Courses, Instructional Materials and Equipment, Instructional and Learning Procedures and Administrative Arrangements," Chapter X, Educational Evaluation, National Society for the Study of Education, University of Chicago, Press, Chicago, July, 1968.

(3) Markle, D. G., "On the Control of Runaway Programmers," close approximation to a speech given at 1967 NSPI Convention, Boston.

(4) Scott, R. O., As Interviewed by Stephen Yelon, July, 1972.

(5) Ibid.

(6) Miller, J. G., "Toward a General Theory for the Behavioral Sciences," The American Psychologist, vol. 10, no. 9 (September, 1955).

(7) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(8) Atkin, J. M., "Some Evaluation Problems in a Course Content Improvement Project," Journal of Research in Science Training," vol. 1, pp. 129-132 (1963).

(9) Smith, R. G., Controlling the Quality of Training, Human Resources Research Office, Alexandria, Virginia, June, 1965.

(10) Briggs, J., Handbook of Procedures for the Design of Instruction, Monograph No. 4, American Institutes for Research, Pittsburgh, 1970.

(11) Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L., Unobtrusive Measures: Nonreactive Research in the Social Sciences, Rand McNally & Co., Chicago, 1966.

(12) Bernstein, L., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1973.

(13) Lindvall, C. M., Cox, R. C., with the collaboration of J. O. Bolvin, Evaluation as a Tool in Curriculum Development: The IPI Evaluation Program, Rand McNally & Co., Chicago, 1970.

(14) Briggs, op.cit.

(15) Ibid.

(16) Schalock, H. Del, "Measurement," Section V, National Research Training Manual, Edited by Jack Crawford, Teaching Research, Division of the Oregon State System of Higher Education, Monmouth, 1969.

(17) Ibid.

(18) Ibid.

(19) Shoemaker, D. M. [illegible]luating the Effectivness of Competing Instructional Programs," Southwest [illegible]nal Laboratory for Educational Research and Development, Los Alamitos, May, 1972.

(20) Review of Educational Research: Educational Evaluation, vol. 40, no. 2 (April, 1970).

(21) Scriven, M., "The Methodology of Evaluation," Perspectives of Curriculum Evaluation, Rand McNally & Co., Chicago, 1967.

(22) Taber, J. I., Glaser, R., and Schaefer, H. N., Learning and Programmed Instruction, Addison-Wesley, Reading, 1965.

(23) Pipe, P., Practical Programming, Holt, Rinehart & Winston, New York, 1966.

(24) Paulson, C. F., "Evaluation of Instructional Systems," Section IV, National Research Training Manual, Edited by Jack Crawford, Teaching Research, Division of the Oregon State System of Higher Education, Monmouth, 1969.

(25) Schutz, R. E., "Experimentation Relating to Formative Evaluation," Research and Development Strategies in Theory Refinement and Educational Improvement, Center for Cognitive Learning, University of Wisconsin Press, Madison, 1967.

(26) Markle; S. M., "Empirical Testing of Programs," Programmed Instruction, Sixty-sixth Yearbook of the National Society for the Study of Education, Part II, Edited by Phil C. Lange, University of Chicago Press, Chicago, 1967.

(27) Briggs, op.cit.

(28) Markle, D. G., The Development of the Bell System First Aid and Personal Safety Course, American Institutes for Research, Palo Alto, 1967.

(29) Anderson, R. C., "The Comparative Field Experiment: An Illustration from High School Biology," Educational Testing Service, Princeton, 1968.

(30) Short, J. G., et. al., "Strategies of Training Development," Report No. AIR - E97 - 2/68, American Institutes for Research, Palo Alto, 1968.

Chapter XIII

(1) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(2) Paulson, C. F., "Evaluation of Instructional System," Section IV, National Research Training Manual, Edited by Jack Crawford, Training Research, Division of the Oregon State System of Higher Education, Monmouth, 1969.

(3) Briggs, L. J., Handbook of Procedures for the Design of Instruction, Monograph No. 4, American Institutes for Research, Pittsburgh, 1970.

(4) Markle, D. G., "On the Control of Runaway Programmers," close approximation to a speech given at 1967 NSPI Convention, Boston.

(5) Taylor, W., "'Cloze Procedures': A New Tool for Measuring Readability," Journalism Quarterly, vol. 30, pp. 415-433 (Fall, 1953).

(6) Baker, R. L., Schutz, R. E., Instructional Product Development, Van Nostrand Reinhold Company, New York, 1971.

(7) Abedor, op.cit.

(8) Palmer, E., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1973.

(9) Abedor, A. J., Bell, N., As Interviewed by Stephen Yelon, Michigan State University, East Lansing, 1973.

(10) Twyford, L., "Profile Techniques for Program Analysis," AV Communication Review, vol. 2, pp. 243-256 (Winter, 1954).

(11) Audience Studies, Inc., A Synopsis of Advertising Testing Services and Procedures, Audience Services, Inc., New York, 1970.

(12) May, M. A., Lumsdaine, A. A., Learning From Films, Yale University Press, New Haven, Conn., 1958.

(13) Weisgerber, R. A., "A Pattern for University Film Production," Journal of the Society of Motion Picture and Television Engineers, vol. 72, pp. 290-291 (April, 1963).

(14) Gliessman, D., Williams, D., "Stimulus Film: Films Made From Pretested Scripts as a Medium for Teacher Education," Audiovisual Instruction, pp. 552-554 (September, 1966).

(15) Knaup, H., Tousey, R., "Producing Test Commercials -- Second-Guessing the First Time Around," The Association of National Advertisers, Inc., New York City, October, 1972, p. 9.

Chapter XIV

(1) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(2) Handel, L. A., Hollywood Looks at Its Audience, University of Illinois Press, Urbana, Ill., 1950.

(3) Review of Educational Research: Educational Evaluation, vol. 40, no. 2 (April, 1970).

(4) Twyford, L., Film Profiles, Technical Report No. SDC 269-7-23, U.S. Naval Training Devices Center, Port Washington, November, 1951.

(5) Popham, J. W., Baker, E. L., Rules for the Development of Instructional Products, Southwest Regional Laboratory for Educational Research and Development, Inglewood, 1967.

(6) Rose, N., Van Horn, C., "Theory and Application of Preproduction Testing," AV Communication Review, vol. 4, pp. 21-30 (Winter, 1956).

(7) Husek, T. R., Sviotnik, K., Item Sampling in Educational Research, Educational Resources Information Center, Washington, D.C., 1967.

(8) Horn, R. E., Development Testing, Center for Programmed Learning for Business, Ann Arbor, 1966.

(9) Markle, S. M., "Empirical Testing of Programs: Dimension, Problems and Issues," Unpublished draft for Sixty-sixth Yearbook of the National Society for the Study of Education, Chicago, March-April, 1965.

(10) Ibid.

(11) Rosen, M. J., An Experimental Design for Comparing the Effects of Instructional Media Programming Procedures, Ed.D. Thesis, University of California, Los Angeles, 1968.

(12) Silberman, H., Coulson, J., Use of Exploratory Research and Individual Tutoring Techniques for the Development of Programming Methods and Theory, System Development Corp., Santa Monica, June, 1965.

(13) Light, J. A., "Formative Evaluation Procedures for the In-Context Development of Instructional Materials," Annual Meeting of the AERA, Chicago, 1972.

(14) Nimnicht, G. P., Rayder, N. F., and Alward, K., An Evaluation of Nine Toys and Accompanying Learning Episodes in the Responsive Model Parent/Child Component, Occasional Research Report No. 5, Far West Laboratory for Educational Research and Development, Berkeley, June, 1970.

Chapter XV

(1) Chen, M., "Verbal Response to 'The Electric Company': Qualities of Program Material and the Viewing Conditions Which Affect Verbalization," Children's Television Workshop, New York, 1972.

(2) Grass, R. C., "Prediction of Television Commercial Field Performance Using Laboratory Techniques," The Association of National Advertisers, Inc., New York, October. 1972.

(3) Scott, R. O., "Program Development Through Formative Evaluation: The SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, August, 1970.

(4) Bryant, Jr., J., "Measures of Attention for C.T.W. Programs," Children's Television Workshop Memorandum, New York, 1972.

(5) Anderson, R. C., "The Comparative Field Experiment: An Illustration from High School Biology," Educational Testing Service, Princeton, 1968.

(6) Rayder, N. P., As Interviewed by Stephen Yelon, Far West Laboratory for Educational Research and Development, Berkeley, January, 1973.

Chapter XVI

(1) Stake, R. E., "Language, Rationality, and Assessment," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, National Education Association, Washington, D.C., 1969.

(2) Scheier, E., Kormac, L. M., and Senter, D. R., A Summary of the Evaluation of the Educational Developmental Laboratories/American Institute of Banking High School Equivalency Program for Bank Trainees, Report No. 5, Education Development Laboratory, New York, March, 1972.

(3) Stufflebeam, D. L., "Evaluation as Enlightenment for Decision Making," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, National Education Association, Washington, D.C., 1969.

(4) Dayton, C. M., "Implications of Educational Research of the Phenomenon of Experimenter Bias," Educational Leadership, vol. 24, pp. 733-739 (1967).

(5) Fanning, J. F., "Implications of Overt Manifestations of Expectancy Bias," Educational Leadership, vol. 25, pp. 683-687 (1968).

(7) Light, J. A., "Formative Evaluation Procedures for the In-Context Development of Instructional Materials," Annual Meeting of the AERA, Chicago, 1972.

(8) Hamreus, D. G., "Instructional Systems Development," Section III, *National Research Training Manual*, Edited by Jack Crawford, Training Research, Division of Oregon State System of Higher Education, Monmouth, 1969.

(9) Westbury, I., "Curriculum Evaluation," *Review of Educational Research: Educational Evaluation*, vol. 40, no. 2 (April, 1970).

(10) Salomon, G., Eglstein, S., Finkelstein, R., Finkelstein, I., Mintzberg, E., Malve, D. and Valner, L., "Educational Effects of 'Sesame Street' on Israeli Children (Brief Summary)," The Hebrew University of Jerusalem, Jerusalem, September, 1972.

(11) Locatis, C., Smith, F., "Guidelines for Developing Instructional Products," *Educational Technology* (April, 1972).

(12) Hamreus, *op.cit.*

(13) Paulson, C. F., "Evaluation of Instructional Systems," Section IV, *National Research Training Manual*, Edited by Jack Crawford, Teaching Research Division of Oregon State System of Higher Education, Monmouth, 1969.

(14) Abedor, A. J., *Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems*, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(15) Sanders, J. R., "Comments on Four Case Studies of Formative Evaluation," Educational Research and Evaluation Laboratory, Indiana University, Bloomington, September, 1971.

(16) Paulson, *op.cit.*

(17) Abedor, *op.cit.*

(18) Paulson, *op.cit.*

(19) Abedor, *op.cit.*

(20) Rose, N., Van Horn, C., "Theory and Application of Pre-Production Testing," *AV Communication Review*, vol. 4, pp. 21-30 (Winter, 1956).

(21) Abedor, *op.cit.*

(22) Reeves, B., "Memorandum: Report of Research on Five Test Shows," Children's Television Workshop, New York, 1969.

(23) Lindvall, C. M., Cox, R. C., *Evaluation as a Tool in Curriculum Development: The IPI Evaluation Program*, Rand McNally & Co., Chicago, 1970.

(24) Audience Studies, Inc., *A Synopsis of Advertising Testing Services and Procedures*, Audience Studies, Inc., New York, 1970.

(25) Maloney, J., As Interviewed by Stephen Yelon, Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, January, 1973.

(26) Goode, H. H., Machal, R. E., *System Engineering - An Introduction to One Design of Large-Scale Systems*, McGraw-Hill, New York, 1957.

(27) Klein, S., As Interviewed by Stephen Yelon, UCLA, Los Angeles, January, 1973.

(28) *Ibid.*

Chapter XVII

(1) Grass, R. C., "Disciplined Creativity: The Latest Shotgun Wedding," The Association for National Advertisers, Inc., New York, February 29, 1968, p. 4.

(2) Nimnicht, G. P., "A Report on the Development and Evaluation of the Parent/Child Toy-Lending Library Program," Far West Laboratory for Educational Research and Development, Berkeley, California, 1971, p. 14.

(3) Ibid.

(4) Baldwin, J. S., "Evaluation of Learning in Industrial Education," Chapter 23, Handbook on Formative and Summative Evaluation of Student Learning, Bloom, B. S., Hastings, J. T., and Madaus, G. F., McGraw-Hill Book Co., New York, 1971.

(5) Watkins, R., As Interviewed by Stephen Yelon, Far West Laboratory for Educational Research and Development, Berkeley, California, 1973.

(6) Sanders, J. R., "Comments on Four Case Studies of Formative Evaluation," Educational Research and Evaluation Laboratory, Indiana University, Bloomington, September, 1971.

(7) Scott, R. O., "Program Development Through Formative Evaluation: The SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, 1970.

(8) Scott, R. O., Martin, M. F., "The 1969-1970 Classroom Tryout of the SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, June 9, 1970.

(9) Ibid., p. 16.

(10) Ibid., p. 17.

(11) Ibid., p. 18.

(12) Sedlik, J. M., Systems Engineering of Education XIV: Systems Techniques for Pretesting Mediated Instructional Materials, Education and Training Consultants Co., Los Angeles, 1971.

Chapter XIX

(1) Scott, R. O., Martin, M. F., "The 1969-1970 Classroom Tryout of the SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, June 9, 1970.

(2) Ibid.

(3) Lindvall, C. M., Cox, R. C., with the collaboration of John O. Bolvin, Evaluation as a Tool in Curriculum Development: The IPI Evaluation Program, Rand McNally & Co., Chicago, 1970.

(4) Ibid., p. 48.

(5) Ibid., p. 104.

(6) Scott, op.cit.

(7) Anderson, R. C., "The Comparative Field Experiment: An Illustration from High School Biology," Educational Testing Service, Princeton, 1968.

(8) Metfessel, N. S., Michael, W. B., "A Paradigm Involving Multiple Criterion Measures for the Evaluation of the Effectiveness of School Programs," Educational and Psychological Measurement (1967).

(9) Emmer, E. T., Millett, G. B., Improving Teaching Through Experimentation, A Laboratory Approach, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

(10) Paulson, C. F., "Evaluation of Instructional Systems," Section IV, National Research Training Manual, Edited by Jack Crawford, Teaching Research, Division of Oregon State System of Higher Education, Monmouth, 1969.

(11) Sidman, M., Tactics of Scientific Research, Basic Books, Inc., New York, 1960.

(12) Paulson, op.cit.

Chapter XX

(1) Light, J. A., "Formative Evaluation Procedures for the In-Context Development of Instructional Materials," Annual Meeting of the AERA, Chicago, 1972.

(2) Light, J. A., *The Development and Application of a Structured Procedure for the In-Context Evaluation of Instructional Materials*, Master's Thesis, University of Pittsburgh, Pittsburgh, 1972, p. 50.

(3) *Ibid*., pp. 16-17.

(4) Abedor, A. J., *Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems*, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(5) Bimonte, R., "A Creative Man's View of TV Commercials Research," The Association of National Advertisers, Inc., New York, 1972.

(6) Light, J. A., "Formative Evaluation Procedures for the In-Context Development of Instructional Materials," Annual Meeting of the AERA, Chicago, 1972.

(7) Palmer, E., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972.

(8) Fitzpatrick, E. W., "Design of the Multi-Media Economic Analysis Course as a Complete Instructional System," Annual Meeting of the AERA, 1971.

(9) O'Bryan, K., "Memo on Eye-Movement," Children's Television Workshop, New York, July 31, 1972.

(10) Light, J. A., *The Development and Application of a Structured Procedure for the In-Context Evaluation of Instructional Materials*, Master's Thesis, University of Pittsbrugh, Pittsburgh, 1972, p. 50.

(11) *Ibid*., p. 45.

(12) *Ibid*., p. 70.

(13) *Ibid*., p. 77.

(14) Dick, W., "A Methodology for the Formative Evaluation of Instructional Materials," *Journal of Educational Measurement*, vol. 5, pp. 99-102 (1968).

(15) Scott, R. O., Martin, M. F., "The 1969-1970 Classroom Tryout of the SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, June 9, 1970.

Chapter XXI

(1) Rust, L., "Attributes of 'The Electric Company' That Influence Children's Attention to the Television Screen," Children's Television Workshop, New York, 1972.

(2) *Ibid*., pp. 12-13.

(3) *Ibid*., p. 18.

(4) *Ibid*., pp. 24-25.

(5) *Ibid*., p. 35.

Chapter XXII

(1) Baker, R. L., Schutz, R. E., *Instructional Product Development*, Van Nostrand Reinhold Company, New York, 1971.

(2) Kaufman, R. A., "Accountability, A System Approach and the Quantitative Improvement of Education--An Attempted Integration," *Educational Technology,* pp. 21-26 (January, 1971).

(3) Eigen, L., "From Where I Sit," NSPI Journal, p. 5 (July, 1969).

(4) Flanagan, J. C., Jung, S. M., "An Illustration: Evaluating a Comprehensive Educational System," AIR Seminar on Evaluative Research: Strategies and Methods, January, 1970.

(5) Sedlik, M., Systems Engineering of Education XIV: Systems Techniques for Pretesting Mediated Instructional Materials, Education and Training Consultants Co., Los Angeles, 1971.

(6) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(7) Lai, M. K., Gall, M. D., Elder, R. A., and Weathersby, R., Main Field Test Report, Discussing Controversial Issues, Far West Laboratory for Educational Research and Development, Berkeley, 1972.

(8) Scott, R. O., Martin, M. F., "The 1969-1970 Classroom Tryout of the SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, June 9, 1970.

(9) Ibid., pp. 11-12.

(10) Scott, R. O., "Program Development Through Formative Evaluation: The SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, August, 1970, p. 10.

(11) Ibid., p. 12.

(12) Lai, op.cit., p. 68.

(13) Levin, A., "Testing for Communications Effectiveness," The Association of National Advertisers, Inc., New York, 1972.

(14) Lesser, G., "Assumptions Behind the Production and Writing Methods in 'Sesame Street'," in Instructional Television, edited by W. Schramm, The University Press of Hawaii, Honolulu, 1972.

(15) Fortune, T., "Bridges for Poor Readers," Children's Television Workshop Memorandum, New York, May 22, 1973, p. 2.

(16) "The Electric Company" Research Staff, "Comprehension Test Results on Show #206," Children's Television Workshop Memorandum, New York, February 3, 1973, p. 1.

(17) Ibid.

(18) Ibid., p. 2.

(19) Ibid.

(20) Silberman, H., Coulson, J., Use of Exploratory Research and Individual Tutoring Techniques for the Development of Programming Methods and Theory, System Development Corp., Santa Monica, June, 1965.

(21) Ibid., p. 41.

(22) Scott, R. O., "Program Development Through Formative Evaluation: The SWRL Instructional Concepts Program," Southwest Regional Laboratory for Educational Research and Development, Inglewood, 1970, p. 12.

(23) Stone, J., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972.

(24) Rust, L., "Attributes of 'The Electric Company' That Influence Children's Attention to the Television Screen," Children's Television Workshop, New York, 1972.

Chapter XXIII

(1) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(2) Hovland, C. I., Lumsdaine, A. A. and Sheffield, F. D., Experiments on Mass Communication, Princeton University Press, Princeton, N.J., 1949.

(3) Lumsdaine, A. A., "Experimental Research on Instructional Devices and Materials," Training Research in Education, Edited by Robert Glaser, John Wiley and Sons, Inc., New York, 1962.

(4) Shoemaker, D. M., "Evaluating the Effectiveness of Competing Instructional Programs," Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, May, 1972.

(5) Ibid.

(6) Anderson, R. C., "The Comparative Field Experiment: An Illustration from High School Biology," Educational Testing Service, Princeton, 1968.

(7) Light, J. A., The Development and Application of a Structured Procedure for the In-Context Evaluation of Instructional Materials, Master's Thesis, University of Pittsburgh, Pittsburgh, 1972.

(8) Abedor, op.cit.

(9) Vander Meer, A. W., An Investigation of the Improvement of Educational Filmstrips and a Derivation of Principles Relating to the Effectiveness of These Media, Phase II, Revision of Film Strip - The Sun and Its Planets, Educational Resources Information Center, Washington, 1964.

(10) Ibid., p. 16.
(11) Ibid., pp. 18-20.
(12) Ibid., p. 72.
(13) Ibid., p. 73.
(14) Ibid., p. 72.
(15) Ibid., p. 73.

Chapter XXIV

(1) Stake, R. E., "Language, Rationality, and Assessment," Improving Educational Assessment and An Inventory of Measures of Affective Behavior, Association for Supervision and Curriculum Development, Washington, D.C., 1969, p. 34.

(2) Paulson, C. F., "Evaluation of Instructional Systems," Section IV, National Research Training Manual, Edited by Jack Crawford, Training Research, Division of Oregon State System of Higher Education, Monmouth, 1969.

(3) Haney, J. B., Lange, P. C., and Barson, J., "The Heuristic Dimension of Instructional Development," AV Communication Review, vol. 16, no. 4 (Winter, 1968).

(4) Palmer, E., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1973.

(5) Paulson, op.cit.

(6) Grasha, A. F., "Evaluating Teaching: Some Problems," University of Cincinnati, Institute for Research and Training in Higher Education, vol. 4, no. 3 (Spring, 1972).

(7) Mielke, K., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1973.

(8) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(9) Baker, R. L., Schutz, R. E., Instructional Product Development, Van Nostrand Reinhold Company, New York, 1971.

Chapter XXV

(1) Abedor, A. J., Development and Validation of a Model Explicating the Formative Evaluation Process for Multi-Media Self-Instruction Learning Systems, Ph.D. Thesis, Michigan State University, East Lansing, College of Education, 1971.

(2) Haney, J. B., Lange, P. C., and Barson, J., "The Heuristic Dimension of Instructional Development," AV Communication Review, vol. 16, no. 4 (Winter, 1968).

(3) Land, H. W., The Children's Television Workshop, How and Why It Works, Nassau Board of Cooperative Educational Services, Jericho, N.Y., 1972, p. 33.

(4) Haney, op.cit.

(5) Palmer, E., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972.

(6) O'Bryan, K., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1973.

(7) Land, op.cit., p. 57.

(8) Palmer, E., As Interviewed by Stephen Yelon, Children's Television Workshop, New York, 1972.

(9) Land, op.cit., p. 46 (quote from G. Lesser).

(10) Ibid., p. 44 (quote from G. Lesser).

(11) Ibid., p. 45 (quote from G. Lesser).

(12) Ibid., p. 47 (quote from G. Lesser).

(13) Ibid., p. 48 (quote from D. Ogilvie).

(14) Ibid., p. 77 (quote from E. Palmer).

(15) Ibid., p. 77 (quote from E. Palmer).

(16) Ibid., p. 63.

(17) Ibid., p. 35 (quote from J. G. Cooney).

(18) Ibid., p. 32 (quote from J. Moss).

(19) Ibid., p. 57 (quote from D. Connell).

(20) Ibid., p. 74.

(21) Ibid., p. 73

(22) Ibid., p. 61 (quote from G. Lesser).

(23) Ibid., p. 74 (quote from J. Stone).